NIST Special Publication 1276v1

NIST Conference Papers
Fiscal Year 2019
Volume 1:
Engineering Laboratory

Compiled and edited by:
Information Services Office

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1276v1

National Institute of
Standards and Technology
U.S. Department of Commerce

NIST Special Publication 1276v1

NIST Conference Papers Fiscal Year 2019

Volume 1: Engineering Laboratory

Compiled and edited by:
Resources, Access, and Data Team
Information Services Office

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1276v1

April 2022

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

National Institute of Standards and Technology Special Publication 1276v1
Natl. Inst. Stand. Technol. Spec. Publ. 1276v1, 737 pages (April 2022)
CODEN: NSPUE2

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1276v1

Foreword

NIST is committed to the idea that results of federally funded research are a valuable national resource and a strategic asset. To the extent feasible and consistent with law, agency mission, resource constraints, and U.S. national, homeland, and economic security, NIST will promote the deposit of scientific data arising from unclassified research and programs, funded wholly or in part by NIST, except for Standard Reference Data, free of charge in publicly accessible databases. Subject to the same conditions and constraints listed above, NIST also intends to make freely available to the public, in publicly accessible repositories, all peer-reviewed scholarly publications arising from unclassified research and programs funded wholly or in part by NIST.

This Special Publication represents the work of researchers at professional conferences, as reported in Fiscal Year 2019.

More information on public access to NIST research is available at https://www.nist.gov/open.

Key words

NIST conference papers; NIST research; public access to NIST research.

Table of Contents

Environmental Friendly Flame Resistant Coatings for Soft Furnishing

*Yeon Seok Kim, Yu-Chin Li, John Shield, and Rick Davis**

Flammability Reduction Group, Fire Research Division, Engineering Laboratory, National Institute of Standards and Technology, 100 Bureau Dr. MS-8665, Gaithersburg, MD 20899-8655 USA

***Abstract*— Pending flammability regulations and scrutinizing of the environment and health consequences of fire retardants is creating opportunities for novel and "greener" technologies to reduce the flammability of residential furniture. This presentation will discuss bio-based fire resistant coatings applied to flexible polyurethane foam and fire blocking barrier fabrics. The waterborne coatings were fabricated using Layer-by-Layer assembly and an innovative one-step/one-pot process and were constructed from common natural materials. The coatings caused significant reductions in the flammability (e.g., heat release, ignition propensity, and flame spread) of these substrates. When used in full-scale fire tests, the coated foam caused as high as a 75% reduction in peak, total, and average heat release of furniture.**

***Index Terms*—One-pot, flame retardant coating, soft furnishing, sodium polyborate, sodium montmorillonite**

I. Introduction

In the United States, there are more than 366,000 residential fires each year. Annually, these fires cause more than 2,500 civilian fatalities and 13,000 civilian injuries [1]. Though one of the lowest in frequency, fires involving residential furniture and mattresses are responsible for the largest fraction of these fatalities and injuries. To significantly reduce the fire severity of soft furnishings, it is critical to eliminate the flexible polyurethane foam from participating in the fire. However, existing fire retardant technologies are not viable options due to their ineffectiveness, and their banning because of potential environment and health concerns.

In 2009, Grunlan et al. (Texas A&M University) first used LbL to produce a fire retardant coating on fabric [2]. Since then Grunlan et al. has continued to be a pioneer in this area by advancing this technology through the research at Texas A&M University, and collaborating with groups at the National Institute of Standards and Technology (NIST) and Polytechnic University of Torino. Over the last several years, these three research groups have been the epicenter of LbL fabricated fire retardant coatings [3-13]. These research groups have developed FR coatings applied to flexible foam and fabrics, constructed of synthetic and bio-based polymer binders (e.g., polyacrylic acid and chitosan), and have contained a range of fire retardants (e.g., sodium polyphosphate and phytic acid) and protective residue formers/enhancers (e.g., montmorillonite clay and layered double hydroxides). These variations and extensions of the original concept have resulted in more rapidly fabricated and highly fire resistant coatings (e.g., a single step process for fabricating a fire retardant coating on fabric [14]).

In 2011, Tsuyumoto et al. reported using starch and sodium polyborate (SPB) to form a fire resistant coating on poly(ethylene terephthalate) and polypropylene non-woven fabrics,[15,16] and rigid polyurethane foam [17]. They reported that these starch-SPB based coatings were able to take these substrates from a few second flame penetration time to no flame penetration in 12 min. In general, this type of flammability reduction required a coating that added more than 50% to the mass of the substrate and contained 12% to 40% SPB. In 2012, Glenn et al, reported using starch and sodium betonite (a layered silicate) to form fire resistant gel coatings for protecting structures against wild land fires [18]. The coatings were applied to exterior cement board lap siding. These coatings increased the time to reach 200 ℃ on the siding surface (a critical fire metric) by as much as 30 min.

In this study, a one-pot process and the chemical formulations to produce a bio-inspired highly fire resistant coating for flexible polyurethane foam (PUF) are investigated. The coatings were constructed of the polysaccharide binder (starch), sodium polyborate, and a protective residue former/enhancer (MMT). Full-scale chair fire tests were conducted to better understand the actual impact of this FR technology under realistic fire conditions.

II. Experimental section

Materials Potato starch (Bob's Red Mill™) were obtained from a local grocery store. Sodium polyborate was obtained from InCide® Technolgies (SPB, Boron #10). Sodium montmorillonite clay was obtained from Southern Clay Products Inc. (MMT, Sodium Cloisite™). Standard (untreated) polyurethane foam (PUF) was obtained from FXI Inc. (Media, PA) and was stored in a climate controlled room with no direct sunlight exposure.

Coating Process and Characterization The coating solutions were prepared by first making the boron FR solution, then adding MMT, and lastly adding the starch. All depositing and washing solutions were water based and were prepared using water purified from a Nanopure II system (18.2 MΩ•cm, Sybron/Barnstead).

SPB (23%) aqueous solutions were prepared by adding SPB to DI water. The solution was heated (60 ℃) and stirred until the SPB fully dissolved and the reaction to form SPB was complete (30 min). 2 mass % MMT powder was added to the SPB solution. The SPB-MMT solution was stirred for a couple hours. Then, the starch powder (3 mass % of the current total

Kim, Yeon; Li, Yu-Chin; Shields, John; Davis, Rick.
"Environmental Friendly Flame Resistant Coatings for Soft Furnishing."
Paper presented at 26th Annual Conference on Recent Advances in Flame Retardancy of Polymeric Materials 2015, Stamford, CT, United States. May 18, 2015 - May 20, 2015.

mixture) was added to the SPB-MMT solution. Then, the solution was heated (90 ℃) and stirred until the solution formed a gel. Coating began once the solution cooled to ~50 ℃. The foam was squeezed and released several times in the solution, then left to soak. After 2 min of soaking, the excess material was squeezed out of the sample and the sample was dried overnight at 70 ℃ in an air convection oven.

A Zeiss Ultra 60 Field Emission-Scanning Electron Microscope (FE-SEM, Carl Zeiss Inc., Thornwood, NY) was used to acquire surface specimens of the coatings on the PUF under a 5 kV accelerating voltage. All SEM samples were sputter coated with 8 nm of Au/Pd (60 %/40 % by mass) prior to imaging. The elementary compositions of the coating were analyzed using energy dispersive x-ray spectroscopy (XEDS) equipped with FE-SEM under 15 kV accelerating voltage. The same samples were used for both SEM and XEDS analysis.

Flammability Testing Cone calorimetry was conducted according to a standard testing procedure (ASTM E-1354-07) with a dual Cone Calorimeter. The Cone was operated with an incident target flux of 35 kW/m^2 and an exhaust flow of 24 L/s.

The chairs were constructed with four cushions (two small ones for the arms and two large ones for the seat and back cushions) in accordance with California Technical Bulletin 133. [19] All cushions were upholstered with 78% polyethylene/22% polyester or 100% cotton cover fabrics. The cushions were assembled on a steel frame representing a chair. The mockup was ignited using a wand constructed from 0.95 cm diameter stainless steel tubing to apply a 3.50 cm long flame, generated by igniting propane gas, at the center of the cavity between the seat and back cushions for 20 s. Heat flux gauges, a One (1) Megawatt (MW) Fire Product Collector (FPC), and a weighing device were used to obtain measurements of the test assembly during the experiments. The experiments were conducted in the Medium Burn Room (MBR) of the Bureau of Alcohol, Tobacco, Firearms and Explosives Fire Research Laboratory (ATF FRL) located in Beltsville, MD.

III. Results and Discussion

The Cone Calorimeter (Cone) is a commonly used instrument to measure bulk flammability characteristics of materials. The sample is exposed to an external heat flux, which forces the material to undergo pyrolysis. Once sufficient fuel (pyrolysis products) is produced, ignition will occur and the sample will undergo combustion and continue to pyrolyze. Cone data and HRR curves for the starch-based coatings are provided in Figure 1.

All of the starch-based coatings reduced the Cone flammability (PHRR and AHRR) of PUF, but had no impact on the THR value (30 ± 3 MJ/m^2). This indicates the coated PUF was completely consumed during combustion, but created a much smaller sized fire then the standard PUF. The best performing formulation was the 3% starch with 23% SPB, which produced a 75% and 81% reduction in the PHRR and AHRR values. The next best formulation was a group of six formulations. These six formulations gave similar Cone results: an average of 63% reduction in PHRR and 72% reduction in AHRR. These formulations include all those that contained 11.5% SPB and those that contained 5.8% SPB with MMT. For the 5.8% formulations without MMT, the 3% starch performed better than the 1.5% starch (approximately 10% better reduction in PHRR and AHRR), but neither performed as well as the six just discussed. However, the 5.8% formulation improved and became one of these six by incorporation of the MMT. Also the flammability was no longer dependent on the % starch in the formulation. Adding MMT had no impact for the higher SPB concentration.

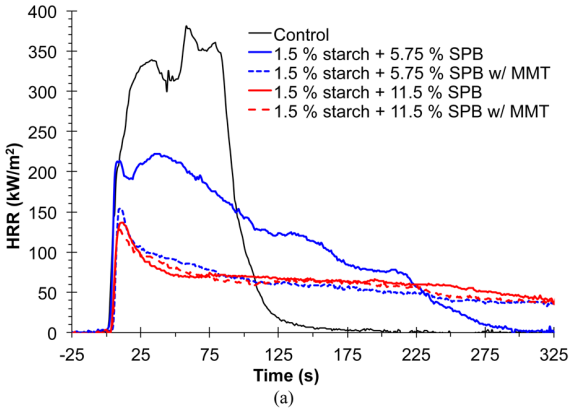


Figure 1. Cone calorimeter heat release rate curves for (a) 1.5% and (b) 3.0% starch-based FR coatings on foam. Uncertainty is ±10% of the reported reduction values. External heat flux was 35 kW/m^2.

The Cone data indicated that SPB itself was sufficient to obtain high fire resistance, but only with a high concentration formulation (e.g., 23%). At lower SPB formulations, a slightly lower flammability was achieved (~10% lower PHRR and AHRR). MMT was needed if the SPB was below a critical threshold (e.g., 5.8%). We decided to conduct full-scale testing on the 1.5% starch-11.5% SBP formulation because we were only slightly compromising flammability in exchange for less raw materials and an easier formulation to coat (higher concentration formulations are more viscous).

Kim, Yeon; Li, Yu-Chin; Shields, John; Davis, Rick.
"Environmental Friendly Flame Resistant Coatings for Soft Furnishing."
Paper presented at 26th Annual Conference on Recent Advances in Flame Retardancy of Polymeric Materials 2015, Stamford, CT, United States. May 18, 2015 - May 20, 2015.

While the Cone is an excellent tool to measure the potential of this fire resistant technology, ultimately the measure of its impact requires full-scale fire tests. The end-use product (e.g., furniture) for this technology is a composite construction (e.g., foam wrapped with fabric and batting) where each component interacts with each other. This interaction can strongly alter the fire behavior. Other factors that influence the fire behavior are the size, shape, and geometry of the product. None of these factors are present in the Cone tests. Therefore, the Cone data was used to quantitatively access the fire resistance of the coatings, where as, the full-scale data was used to quantitatively access the decrease in flammability of furniture built using the fire resistant coated PUF. Full-scale furniture calorimeter data is provided in Figure 2. Time captured images of the full-scale tests are provided in Figure 3.

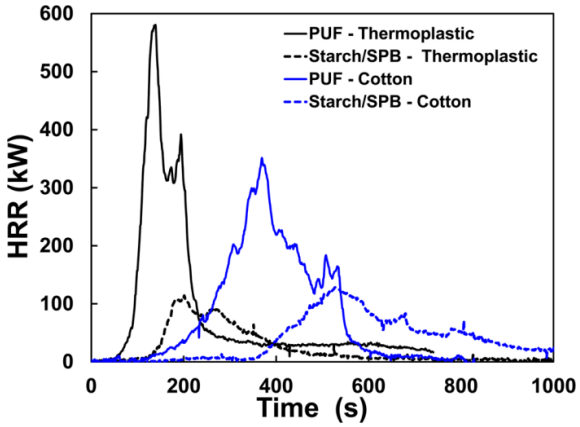


Figure 2. Full-scale fire heat release curves of thermoplastic cover fabric furniture and cotton cover fabric furniture with PUF and 1.5% starch and 11.5% SPB foam. Uncertainty is ±10% of the measured values. External heat flux was 35 kW/m^2.

Full-scale tests indicated the starch-SPB coating might be a better fire retardant technology than suggested by the Cone tests. The back, seat, and arms were all constructed of foam wrapped with a cover fabric. Using a thermoplastic cover fabric and standard PUF, the chair ignited easily and flames rapidly spread across the surface. Within 90 s after ignition, the entire chair was completely engulfed in flames. At 132 s, a PHRR value of 580 kW/m^2 was measured. Less than 2 min later, the test ends with the chair being completely consumed releasing a total heat (THR) of 121 MJ/m^2. The chair was much less flammable by replacing the thermoplastic with a cotton covering fabric. The PHRR was significantly lowered and delayed (350 kW/m^2 at 369 s). The chair was still completely consumed, but released a lower amount of total heat (107 MJ/m^2). This THR difference was due to the thermoplastic releasing more heat than the cotton fabric.

Replacing PUF with the 1.5% starch-11.5% SPB coated foam slowed flame spread, reduced flammability, and caused the furniture to self-extinguish. For the thermoplastic covering fabric chair, at 90 s the PUF chair was completely engulfed in flames whereas the flames still had not spread across the seat of the starch-SPB foam chair. The 71% reduction in THR was due to the starch-SPB foam slowing down pyrolysis to the point that the fuel was insufficient to sustain combustion. Since the thermoplastic cover fabric was completely consumed in both tests, this 71% reduction was directly related to the amount of foam remaining after the test (recall with PUF that the chair was completely consumed). The slower flame spread and lower amount of chair consumed was the reason why this starch-SPB foam resulted in a 75% and 61% reduction in PHRR and AHRR reduction, respectively. For the cotton covering fabric chair, the starch-SPB foam had a similar flammability reduction as observed for the thermoplastic. Normally, the type of covering fabric significantly influences the flammability of a piece of furniture. This was not that case for the starch-SPB coated foam chairs, as the actual test values (except for tPHRR) were independent of the type of covering fabric.

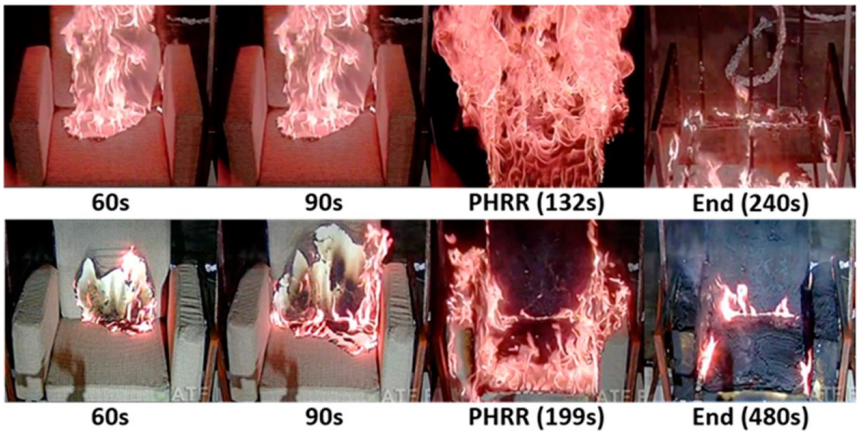


Figure 3. Images from full-scale furniture fire tests with a thermoplastic cover fabric and (top) PUF and (bottom) 1.5% starch and 11.5% SPB foam.

SEM images and XEDS spectrum of a starch-SPB-MMT coating are provided in Figure 4. All the coatings completely encased the foam. There were no features in the SEM images that distinguished one coating from another. All the coatings

Kim, Yeon; Li, Yu-Chin; Shields, John; Davis, Rick.
"Environmental Friendly Flame Resistant Coatings for Soft Furnishing."
Paper presented at 26th Annual Conference on Recent Advances in Flame Retardancy of Polymeric Materials 2015, Stamford, CT, United States. May 18, 2015 - May 20, 2015.

near the outside edge of the foam appeared rough with frequent, large, and flaky aggregates. Near the center of the foam, the coatings appeared significantly smoother with fewer and smaller aggregates. Most of these near the center aggregates appeared to be embedded in the coatings where as these aggregates appeared more as flakes near the edge. Mass of foam samples (1 cm by 1 cm by 1 cm) taken from the edge was on average 20% heavier than from the center. This indicated that the coatings were thicker on the edge, which may explain the rougher and flaky features. We believe the coatings were thicker on the edge because the high viscosity of the depositing solution and the thickness of the coatings significantly slowed down transport into the center of the foam.

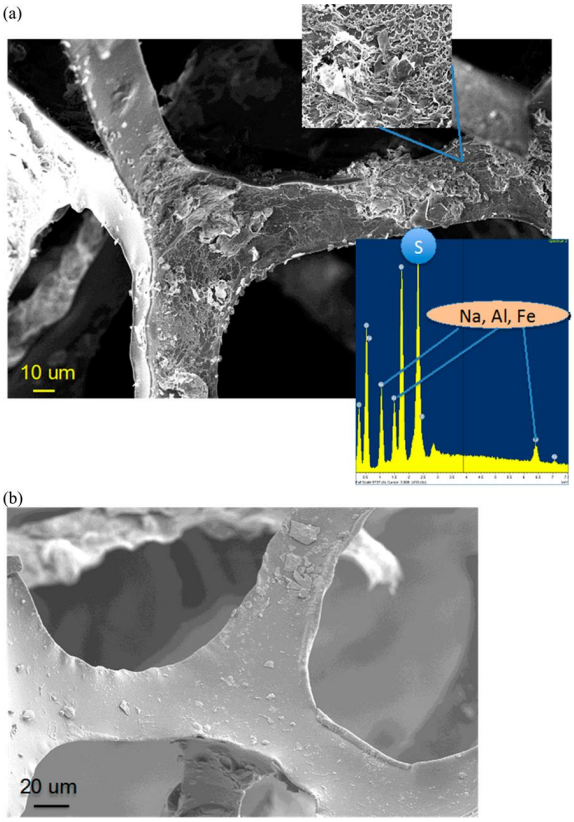


Figure 4. SEM image of 1.5% starch−11.5% SPB−2% MMT coating at the (a) edge and (b) center of the foam. Coating was thicker and has larger flakes and aggregates near the edge. XEDS insert in a shows the presence of S and Na, Al, and Fe in the coatings, which indicate the presence of SPB and MMT. Other peaks are associated with the coating and/or foam.

XEDS was used to determine the presence of MMT and SPB in the coatings. Detecting sodium (Na), magnesium (Mg), iron (Fe), aluminum (Al), and/or silicon (Si) indicated the presence of MMT. Detecting sulfur (S) indicated the presence of SPB. Since boron (B) cannot be resolved from carbon (C), boron could not be detected in any of the coatings. Therefore, there was no unique element that could be used to detect the presence of STB and boric acid. XEDS analysis of a starch-SPB-MMT coated PUF showed the coating contained SPB (S) and MMT (Na, Al, Fe). These elements were detected in all formulations containing SPB and MMT, which indicated these compounds were in the coatings.

IV. Conclusion

Polysaccharide-based coatings applied in a one-step process significantly reduced the flammability of flexible polyurethane foam. The fire resistant coatings were constructed of a starch, sodium polyborate, and/or a MMT. The best performing formulation was a 3% starch-23% SPB, which produced a 75% reduction in the PHRR (as compared to PUF). The effectiveness of this coating technology was validated in full-scale fire tests. Full-scale fire tests of furniture containing a 1.5% starch-11.5% SPB coating produced a 75% lower PHRR than when a standard flexible foam was used. The actual PHRR values were approximately 120 kW/m^2 for the starch-SPB foam chairs as compared to the 580 kW/m^2 and 350 kW/m^2 for the standard PUF chairs. Estimates suggested that the furniture PHRR reduction caused by the starch-SPB coating could reduce the fire threat from potential death and rapid fire spread to low risk of injury and the fire being contained near the burning furniture. To the best of our knowledge, these coatings produced the largest flammability reduction of furniture reported for any fire retarding technology on/in flexible PUF.

Future work has already begun to investigate one-pot fire resistant coatings constructed from other binders, FR, and char forming compounds.

Acknowledgment

The authors would like to thank Dr. William Pitts (NIST) for conducting the full-scale burning experiments.

References

[1] M. Ahrens, Home Structure Fires, National Fire Protection Association, Quincy, MA, 2013.

[2] Y.-C. Li, J. Schulz, J.C. Grunlan, "Polyelectrolyte/Nanosilicate Thin-Film Assemblies: Influence of pH on Growth, Mechanical Behavior, and Flammability," ACS Appl Mater & Interfaces, vol. 1, pp. 2338-2347, 2009.

[3] Y.-C. Li, J. Schulz, S. Mannen, C. Delhom, B. Condon, S. Chang, M. Zammarano, J.C. Grunlan, "Flame Retardant Behavior of Polyelectrolyte-Clay Thin Film Assemblies on Cotton Fabric," ACS Nano, vol. 4, pp. 3325-3337, 2010.

[4] Y.-C. Li, S. Mannen, A.B. Morgan, S. Chang, Y.-H. Yang, B. Condon, J.C. Grunlan, "Intumescent All-Polymer Multilayer Nanocoating Capable of Extinguishing Flame on Fabric," Adv Mater, vol. 23, pp. 3926-3931, 2011.

Kim, Yeon; Li, Yu-Chin; Shields, John; Davis, Rick.
"Environmental Friendly Flame Resistant Coatings for Soft Furnishing."
Paper presented at 26th Annual Conference on Recent Advances in Flame Retardancy of Polymeric Materials 2015, Stamford, CT, United States. May 18, 2015 - May 20, 2015.

[5] G. Laufer, C. Kirkland, A.B. Morgan, J.C. Grunlan, "Intumescent Multilayer Nanocoating, Made with Renewable Polyelectrolytes, for Flame-Retardant Cotton," Biomacromolecules, vol. 13, pp. 2843-2848, 2012.
[6] G. Laufer, C. Kirkland, A.B. Morgan, J.C. Grunlan, "Exceptionally Flame Retardant Sulfur-Based Multilayer Nanocoating for Polyurethane Prepared from Aqueous Polyelectrolyte Solutions," ACS Macro Lett, vol. pp. 361-365, 2013.
[7] F. Carosio, G. Laufer, J. Alongi, G. Camino, J.C. Grunlan, "Layer-by-layer assembly of silica-based flame retardant thin film on PET fabric," Polym Degrad Stabil, vol. 96, pp. 745-750, 2011.
[8] F. Carosio, J. Alongi, G. Malucelli, "Layer by Layer ammonium polyphosphate-based coatings for flame retardancy of polyester–cotton blends," Carbohydrate Polymers, vol. 88, pp. 1460-1469, 2012.
[9] F. Carosio, A. Di Blasio, J. Alongi, G. Malucelli, "Green DNA-based flame retardant coatings assembled through Layer by Layer," Polymer, vol. 54, pp. 5148-5153, 2013.
[10] Y.S. Kim, R. Davis, A.A. Cain, J.C. Grunlan, "Development of layer-by-layer assembled carbon nanofiber-filled coatings to reduce polyurethane foam flammability," Polymer, vol. 52, pp. 2847-2855, 2011.
[11] Y.-C. Li, Y.S. Kim, J. Shields, R. Davis, "Controlling polyurethane foam flammability and mechanical behaviour by tailoring the composition of clay-based multilayer nanocoatings," J Mater Chem A, vol. 1, pp. 12987-12997, 2013.
[12] Y.S. Kim, Y.-C. Li, W.M. Pitts, M. Werrel, R.D. Davis, "Rapid Growing Clay Coatings to Reduce the Fire Threat of Furniture," ACS Appl. Mater. Inter., vol. 6, pp. 2146-2152, 2014.
[13] Y.-C. Li, Y.H. Yang, J.R. Shields, R. Davis, "Layered Double Hydroxide-Based Fire Resistant Coatings for Flexible Polyurethane Foam," Submitted to Polymer, vol. pp. 2014.
[14] A.A. Cain, S. Murray, K.M. Holder, C.R. Nolen, J.C. Grunlan, "Intumescent Nanocoating Extinguishes Flame on Fabric Using Aqueous Polyelectrolyte Complex Deposited in Single Step," Macromol Mater Eng, vol. pp. n/a-n/a, 2014.
[15] I. Tsuyumoto, Y. Miura, M. Nirei, S. Ikurumi, T. Kumagai, "Highly flame retardant coating consisting of starch and amorphous sodium polyborate," J Mater Sci, vol. 46, pp. 5371-5377, 2011.
[16] I. Tsuyumoto, Y. Miura, Y. Hori, "Fire-resistant nonwovens of EVOH and PET treated with amorphous sodium polyborate," J Mater Sci, vol. 45, pp. 2504-2509, 2010.
[17] I. Tsuyumoto, Y. Onoda, F. Hashizume, E. Kinpara, "Flame-retardant rigid polyurethane foams prepared with amorphous sodium polyborate," J Appl Polym Sci, vol. 122, pp. 1707-1711, 2011.
[18] G.M. Glenn, G. Bingol, B.-S. Chiou, A.P. Klamczynski, Z. Pan, "Sodium bentonite-based coatings containing starch for protecting structures in wildfire emergency situations," Fire Safety J, vol. 51, pp. 85-92, 2012.
[19] California TB 133 Flammability Test Procedure for Seating Furniture for Use in Public Occupancies, 1991.

Kim, Yeon; Li, Yu-Chin; Shields, John; Davis, Rick.
"Environmental Friendly Flame Resistant Coatings for Soft Furnishing."
Paper presented at 26th Annual Conference on Recent Advances in Flame Retardancy of Polymeric Materials 2015, Stamford, CT, United States. May 18, 2015 - May 20, 2015.

Networked Engineering Notebooks for Smart Manufacturing

Peter Denno,[a,1] Charles Dickerson[b], and Jennifer Harding[b]
[a] *National Institute of Standards and Technology, United States*
[b] *Loughborough University*

Abstract. A goal of the industrial internet is to make information about manufacturing processes and resources available wherever decision making may be required. Agile use of information is a cornerstone of data analytics, but analytical methods more generally, including model-based investigations of manufacturability and operations, do not so easily benefit from this data. Rather than relating anonymous patterns of data to outcomes, these latter analytical methods are distinguished as relying on conceptual or physics-based models of the real world. Such models require careful consideration of the fitness of the data to the purpose of the analysis. Verification of these analyses, then, is a significant bottleneck. A related problem, that of ascertaining reproducible results in scientific claims, is being addressed through executable notebook technology. This paper proposes to use notebook technologies to address that bottleneck. It describes how this notebook technology, linked to internet-addressable ontologies and analytical metamodels, can be used to make model-based analytical methods more verifiable, and thus more effective for manufacturers.

Keywords. Manufacturing analysis, process analysis, process optimization, analytical methods, empirical methods, industrial internet of things, metamodels

1. Introduction

Smart manufacturing [1] and similar initiatives [2], [3] are focused on improving manufacturing productivity through the use of inexpensive sensor networks, information systems, and software analytical tools. In manufacturing, a great variety of software analytical tools are being applied to an ever-expanding set of physical [4] and operational [5] problems including process optimization. However, in manufacturing, more so than engineering design, the preparation of analytical models and the interpretation of their results do not share a common methodology across the various usages. This limits the ability to develop systematic means to apply analytical techniques to decision making. The consequences of this limitation include missed opportunities to reuse knowledge, unreliable results, and high cost of analysis. A report from the NIST [6] suggests that it should be possible, near-term, to reduce the cost of system verification and validation ten-fold. But how will this reduction be achieved?

There is no silver bullet to making analytical techniques more accessible to more manufacturers, but opportunities have emerged with the development of technologies that support 1) the industrial internet of things, 2) manufacturing domain ontologies, 3) domain specific languages, 4) natural language processing, and 5) metamodeling. This paper describes how these technologies may be applied using the Python programming language and Jupyter Notebooks which are web applications for the creation and sharing of documents that embody analyses. Notebooks contain live code, equations,

[1] Corresponding Author: Peter Denno, email: peter.denno@nist.gov.

Denno, Peter; Christopher, Chang.
"Networked Engineering Notebooks for Smart Manufacturing."
Paper presented at 14th Annual Conference on Systems Engineering Research (CSER 2016), Huntsville, AL, United States. March 22, 2016 - March 24, 2016.

visualizations and explanatory text [7]. The paper describes a notebook-based methodology, tailored for use in smart manufacturing environments, to reduce the cost of developing and validating process optimizations.

Section 2 of the paper describes the problem space. Section 3 explains how "equations as objects," formal requirements specifications, notebook annotations, and a metamodel of optimization are used in our methodology to produce analyses that are verified and integrated into manufacturing operations. Ideas are illustrated using a case study implementing a turning operation optimization described by Abdelmaguid and El-Hossany [8]. Section 4 concludes the paper with a discussion of limitations of the described process and planned future work.

2. From Analytical Problems to Systematic Solutions

Information needed to formulate manufacturing process optimizations - for example, to improve machining operations - may originate from many viewpoints. A viewpoint is a set of related concerns relevant to a particular audience [9]. Decision making in manufacturing typically relies on a composition of such viewpoints. For example, deciding what jobs to initiate in a job-shop environment might involve a composition of viewpoints concerning product demand, process plans, machine capabilities, machine instances and their availability, and inventory. The problem of composing viewpoints for decision making (i.e. of formulating the relevant analysis) is, in part, the problem of knowing what can be inferred from the union of information sources. This fundamental problem resists simple solution [10].

This paper describes a method in which it is assumed that the most creative part of the process, the problem formulation, has already been performed (perhaps guided by analyses described in the literature [8]). The practical problem that remains is that of making the analysis correct and effective for the idiosyncrasies of the given manufacturing context and the supporting information technology.

Verifying the correctness of an analysis requires knowing 1) what the analysis is to achieve (i.e. its requirements), 2) that the method (algorithm) chosen is suitable to address these requirements, and 3) that inputs to the algorithm are appropriate to it. With this knowledge, one gains confidence in the validity of the analysis, but not an understanding of the bounds on certainty, the sensitivity of parameters, nor the risk that the implied recommendations entail. Of these four qualities (validity, certainty, sensitivity and risk), validity is the most basic requirement. Though the others are important, they are not discussed in this paper.

3. Verifying Process Optimizations in Smart Manufacturing Environments

The literature contains an abundance of mathematical and physics-based models of manufacturing processes. Some of these models are "simple", since they are neither computationally demanding, nor require nuanced understanding of model parameters. Simple models can be applied more directly in daily production operations. For example, such models can be used by manufacturing engineers to adjust operating parameters in reaction to variation in raw materials.

Other models are "complex" since they require extensive setup or a geometric model (such as is the case with finite element analysis, for example). Abstractions can

Denno, Peter; Christopher, Chang.
"Networked Engineering Notebooks for Smart Manufacturing."
Paper presented at 14th Annual Conference on Systems Engineering Research (CSER 2016), Huntsville, AL, United States. March 22, 2016 - March 24, 2016.

sometimes be found over complex models so as to create a simple model more suitable for daily use. An example of such an abstraction is a reduced-order surrogate model based on computational experiments performed on a complex model [11]. This paper concerns simple models and the simple form of complex models.

The methodology requires three classes of information to verify the correctness of an analysis. First is information linking variables and relations to terms in an ontology defining the intended meaning, dimensionality, and sources for values (e.g. a schema providing dimensional inspection results). Second is information linking calls to specialized analytical tools (e.g. optimizers) to a conceptual model of that tool. Third is information containing formal statements of requirements describing what the analysis is to achieve. These three elements are discussed, in turn, in the following sections.

Verification is performed against analyses encoded in Jupyter notebooks using the Ipython language. Ipython is the interactive version of the Python language. Both Ipython and Python can use the many libraries developed for the language. Of particular value to manufacturing analysis are the scipy and numpy libraries for general mathematical tasks and optimization, the Pandas library for data structures and data analysis, and Statsmodels for statistics [12].

Jupyter represents user content (i.e. the analysis) as a Javascript Object Notation (JSON) data structure. The data structure (and its presentation in the browser) is divided into cells. A cell can be designated as containing (Python) code or markdown syntax annotations. Markdown cells can use LaTeX notation to represent mathematical formulas.

The design of the verification method is such that it does not interfere with the execution of the analysis. If the analysis provides markdown cells containing tables describing variables, these can be used in a separate process to annotate the analysis. If, on the other hand, such tables are not provided and sufficient information cannot be inferred through other means, verification will be incomplete.

3.1. Linking Variables to Ontology Terms and Sources

The JSON data structure can be parsed to identify variables defined in tables provided by the user; as is typical of Jupyter notebooks, these tables are in markdown syntax (See Figure 1). If the tables have the appropriate form, they will be interpreted as containing definitions of variables used in the analysis. Additionally, the process managing the validation can enable the user to associate OWL ontology terms with these variables. The ontology provides definitions and constraints on interpretation of elements important to the analysis such as symbols, variables, and their bindings to definitions. These definitions may concern manufacturing concepts.

Symbol	Variable	Meaning
D	D	Final product target diameter
D_0	d_i	depth of cut for part *i*
δ_i	delta_i	tool wear compensation (mm) for part *i*
V	V	Cutting speed (m/min)
f	f	Feed rate (mm/rev)
N	N	Tool regrind scenario

```
| Symbol     | Variable | Meaning |
|-------------------------------------------------------------|
| $D$        | D        | Final product target diameter|
| $D_0$      | d_i      |depth of cut for part *i* |
| $\delta_i$ | delta_i |tool wear compensation (mm) for part *i*|
| $V$        | V        |Cutting speed (m/min)|
| $f$        | f        | Feed rate (mm/rev) |
| $N$        | N        |Tool regrind scenario |
```

Figure 1. A table as it appears in the Jupyter notebook (left), and the corresponding markdown syntax provide by the user (right).

Similarly, the analysis can be annotated with knowledge of the sources of values (e.g. data conforming to XML schema or the schema translated to an OWL ontology).

Software has been written to parse the LaTeX-syntax mathematics in markdown cells to OWL statements. Because mathematical expressions can have complex syntax, a single mathematical expression can produce many interrelated OWL facts. This makes processing difficult, but not impossible. The software implements an "equations as objects" method of operation described in prior work [13]. By making mathematical relations visible outside of the tools in which they are initially developed, these relations can be reused and refined. For example, an empirically-defined, predictive model of a manufacturing process developed in one analysis (one notebook) can be referenced in an optimization or linked to additional operational data and refined in other notebooks.

3.2. Characterizing the Use of Optimizers with an Optimization Metamodel

A principal challenge in using analytical tools such as optimizers is that of knowing what pattern of tool use is appropriate to the problem at hand. There are two aspects to addressing this challenge: representing the domain problem and representing the capabilities and interface of the analytical tool. In the method, the domain problem is represented through a combination of the ontology links discussed in the previous section and the formal requirements characterization discussed in the next section. This section discusses how the user's specification of the objective function, constraints and call to the optimizer are characterized for subsequent validation.

The characteristics of optimization techniques can be modeled using a metamodel. A *metamodel,* as used here, is a model of a modeling language providing sufficient detail to serve as the storage form for instances of the model. Prior work [14] has produced a preliminary UML-based metamodel of optimization to represent optimization problems defined in the Optimization Programming Language (OPL). This metamodel is still under development and being adapted to represent the optimization techniques found in Python scipy. Depicted in Figure 2 is an optimization call such as used in the turning optimization. The vector *x* is the design vector; *x*[0] the cutting speed, and *x*[1] the feed rate. As shown, these are initialized to [60.0, 0.08] respectively. For brevity, only part of the constraint vector is shown.

```
cons =  ({'type': 'ineq',
          'fun' : lambda x: np.array([wear_limit_calc(N,x[0],x[1])
                                      - W_hat])})

def obj_func (x,sign=-1):
    return sign*(P_calc(x[0],x[1],N)/(N*t_h + T_n_calc(x[0],x[1],N)))

res = minimize(obj_func, [60.0,0.08], bounds=[(V_min, V_max), (f_min, f_max)],
               constraints=cons,
               method='SLSQP', options={'disp': True})
```

Figure 2. The call to the optimizer in the turning example.

The optimization metamodel describes a conceptualization of optimization problems. The use of the optimizer, such as in the code in Figure 2, represents a particular instance of the metamodel. This instance can be expressed as RDF and OWL triples. In creating this instance, it is essential to recognize how the information provided in the call to the optimization tool relates to the optimization metamodel. For example, in the case of the Scipy-based tool depicted in Figure 2, the parameter named *method* indicates the

algorithm applied (sequential least squares programming); *bounds* provides constants bounding the feasible region; and, *constraints* provides a vector of functions that can be called with the design vector to evaluate inequality and equality constraints. Integrating this knowledge with variable and equation definitions acquired through the markdown cells requires parsing the call to *minimize* and related constraint definitions.

3.3. Stating Requirements about Analyses

Verification is performed against requirements that originate as informal statements about what the system (in our case an analysis) is to do. Prior work [15] provides a method to construct a controlled, natural-language parser for a domain-specific language. The method involves recognizing what the sentence intends by first classifying its verb by calculating the semantic distance to it from synonyms of all "proto-verbs" of the domain-specific language. The current implementation of the method does not provide such a parser; it only provides similar resulting statements based on a pilot set of proto-verbs: "optimize" "constrain" and "find." This set is likely to grow; the prior work required 19 proto-verbs. The resulting statements are encoded in a dialect of ISO Common Logic. An example statement in the dialect – one describing the requirement that the analysis should maximize profit under constraint of acceptable surface finish and dimension tolerance – is as follows:

```
(obligation-of
   analysis-x
   (that
      (and (optimize profit)
           (under-constraint surface-finish)
           (under-constraint dimensional-tolerance))))
```

3.4. Using Links, Metamodels, and Requirements to Verify Analyses

Inconsistencies in the analysis are identified using a Bayesian approach on a graph-based structure. The process is similar to that described by Herzig [16]. In the process, a graph of binary relations is produced from the sources described in the previous three sections (ontology-annotated variables and terms, the optimization metamodel instance, and requirement statements).

As an example of how inconsistencies are recognized, consider the segment of the turning analysis depicted in Figure 2. *W_hat* in the figure is defined through links to the ontology to be the maximum acceptable tool wear. The translation of the Python constraint through the optimization metamodel indicates that the expression of the constraint on tool wear, `wear_limit_calc(N, x[0], x[1], ) - W_hat`, has its sign reversed from what is correct. (Inequality constraints provided to the python minimization function should be stated such that their values are positive inside the feasible region.) The fact that tool wear is generally not a desirable quality of machining and yet the analysis seemingly seeks a value of tool wear *above W_hat* allows Bayesian inference to identify the inconsistency against the world model reflected by the ontology.

4. Conclusion

This paper discusses a methodology for performing and verifying optimization-based analyses of manufacturing processes and operations. The method applies a "soft system perspective" [2] to facilitate the use of analytical tools in smart manufacturing environments. Elements of the method have been implemented on the turning optimization described in [8] as well as an empirical model of an additive manufacturing process [13]. Much work remains in developing the analytical metamodel for optimization and integrating system components. An analysis for job shop scheduling is being developed. As production scheduling can involve very different viewpoints from those used in unit manufacturing processes, it will be necessary to develop the ontology towards these viewpoints.[3]

References

[1] J. Davis, T. Edgar, J. Porter, J. Bernaden, and M. Sarli, "Smart Manufacturing , Manufacturing Intelligence and Demand-Dynamic Performance," in *FOCAPO2012: Foundations of Computer-Aided Process Operations*, 2012.

[2] VDE-DKE, "The German Standardization Roadmap Industrie 4.0," *Vde Assoc. Electr. Electron. Inf. Technol.*, vol. 0, pp. 1–60, 2014.

[3] Government Office for Science, "The Future of Manufacturing," London, 2013.

[4] K. Allen, Dell, Alting, Leo, and H. Todd, Robert, *Fundamental Principles of Manufacturing Processes*. Industrial Press, 2005.

[5] J. Li and S. M. Meerkov, *Production System Engineering*. Springer Science+Business Media, 2009.

[6] National Institute of Standards and Technology, "Foundations for Innovation in Cyber-Physical Systems," Gaithersburg, 2013.

[7] Project Jupyter, "Jupyter Notebooks." [Online]. Available: http://jupyter.org/. [Accessed: 28-Apr-2016].

[8] T. F. Abdelmaguid and T. M. El-hossainy, "Optimal Cutting Parameters for Turning Operations with Costs of Quality and Tool Wear Compensation," *Proc. 2012 Int. Conf. Ind. Eng. Oper. Manag. Istanbul, Turkey, July 3 – 6*, pp. 924–932, 2012.

[9] ISO, *ISO/IEC/IEEE 42010:2011 Systems and software engineering — Architecture description*. 2011.

[10] N. F. Noy, "Semantic Integration: A Survey Of Ontology-Based Approaches," *Newsl. ACM SIGMOD Rec.*, vol. 33, no. 4, pp. 65–70, 2004.

[11] A. Jeang, H. C. Li, and Y. C. Wang, "A computational simulation approach for optimising process parameters in cutting operations," *Int. J. Comput. Integr. Manuf.*, vol. 23, no. 4, pp. 325–340, 2010.

[12] W. McKinney, *Python for data analysis*. O'Reilly, 2013.

[13] P. O. Denno and D. B. Kim, "Integrating views of properties in models of unit manufacturing processes," *Int. J. Comput. Integr. Manuf.*, no. March, pp. 1–17, 2015.

[14] I. Assouroko and P. O. Denno, "A metamodel for optimization problems," Gaithersburg, 2016.

[15] P. O. Denno and C. Chang, "Validating controlled English statements of requirements using functional models," 2016, pp. 1–10.

[16] S. J. I. Herzig, "A Bayesian Learning Approach To Inconsistency Identification in Model-Based Systems Engineering," Georgia Institute of Technology, 2015.

[17] P. Checkland, "Soft Systems Methodology: A Thirty Year Retrospective," *Syst. Res. Behav. Sci. Syst. Res*, vol. 17, pp. 11–58, 2000.

[2] Checkland [17] distinguishes "hard system perspectives' from "soft systems perspectives": in the former, the world is viewed as systemic; in the later, the world is viewed as unmanageably complex, but the process of inquiry is systemic.

[3] Certain commercial software products are identified in this paper. These products were used only for demonstration purposes. This use does not imply approval or endorsement by NIST, nor does it imply these products are necessarily the best available for the purpose.

Denno, Peter; Christopher, Chang.
"Networked Engineering Notebooks for Smart Manufacturing."
Paper presented at 14th Annual Conference on Systems Engineering Research (CSER 2016), Huntsville, AL, United States. March 22, 2016 - March 24, 2016.

46th SME North American Manufacturing Research Conference, NAMRC 46, Texas, USA

Implementation of Advanced Laser Control Strategies for Powder Bed Fusion Systems

H. Yeung [a]*, B.M. Lane [a], M. A. Donmez [a], J.C. Fox [a], J. Neira [b]

[a] *Engineering Laboratory,* [b] *Physical Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899*

* Corresponding author. Tel.: +1-301-975-2786
E-mail address: ho.yeung@nist.gov

Abstract

Laser path, scan speed, and laser power are critical machine parameters for determining the quality of the output of laser-based powder bed fusion (LPBF) processes. A jerk-limited control strategy is implemented for laser path planning on a LPBF additive manufacturing (AM) testbed. The actual and commanded laser paths/velocities are found to be in better agreement with each other compared to conventional controls. The new controller enabled implementation of advanced laser power control strategies synchronized with laser position and velocity by embedding all into a modified G-code (referred as AM G-code). An interpreter is developed to utilize sophisticated LPBF laser control commands.

Peer-review under responsibility of the scientific committee of the 4th International Conference on System-Integrated Intelligence.

Keywords: Laser Powder Bed Fusion Additive Manufacturing (LPBF AM), Scan Strategies, AM G-code, Jerk-limited Control

1. Introduction

Laser powder bed fusion (LPBF) is an additive manufacturing (AM) process in which a focused, high power laser selectively melts geometric patterns into layers of metal powder, ultimately building a near fully dense freeform part [1]. The LPBF fabrication process, and the resulting part quality, are influenced by hundreds of controlled and uncontrolled process parameters [2]. To form fully dense parts, laser position, velocity, and power must be well controlled based on the powder layer characteristics (material, density, thickness, etc.) to adequately fuse adjacent scan tracks and previous layers. Improper combination of these parameters can cause defects that plague LPBF parts. Pores, for example, have been attributed to various phenomena related to the laser power-velocity profiles or scan strategies (e.g., keyholing and pore entrapment at high laser energy densities [3]), or insufficient re-melting of adjacent scan tracks, often called 'lack of fusion' [4–6]. Better controlled velocity or power profiles, or power density, along each scan

Yeung, Ho; Lane, Brandon; Donmez, M; Fox, Jason; Neira, Jorge.
"Implementation of Advanced Laser Control Strategies for Powder Bed Fusion Systems."
Paper presented at 46th SME North American Manufacturing Research Conference, NAMRC 46, College Station, TX, United States. June 18, 2018 - June 22, 2018.

path can reduce the probability of defect formation, or provide a parametric space for process optimization [7,8].

The laser control for LPBF systems involves both laser path and laser power. A focused laser spot is directed to the powder bed by a pair of mirrors driven by galvanometer (galvo) motors, therefore laser path control is achieved by controlling the two galvo motors in a coordinated manner. Laser power is electronically adjusted through the laser amplifier, usually by a digital 'gate' signal to turn the laser on/off, and a low-voltage analog signal proportional to laser power. Most commercial scanning systems, both standalone and integrated into LPBF machines, use a step velocity profile for motion control. Step velocity assumes infinite acceleration, making it impossible for the mirrors to truly follow a command. Therefore, temporal and spatial accuracy are compromised leading to geometric inaccuracies and material defects.

The National Institute of Standards and Technology (NIST) is constructing an open architecture Additive Manufacturing Metrology Testbed (AMMT) [9] to study advanced process monitoring and control strategies. The work described here is used to implement the laser control on the AMMT. The AMMT is instrumented with a high-speed camera coaxially aligned with the laser for *in-situ* melt-pool imaging. Laser position and power are measured at 100 kHz from the galvo position feedback and laser source unit, respectively [10]. All experiments in this paper were conducted on the NIST AMMT.

2. Influence of laser control on scan path accuracy

In numerical control of machine tools, a jerk-limited path is usually used to avoid excitation of vibration modes in the mechanical structure [11,12]. A jerk-limited path has a smooth velocity profile which is more easily followed by a physical system, and results in better spatial and temporal path accuracy. Here, spatial accuracy refers to geometric position of the laser spot, and temporal accuracy refers to the spot reaching designated positions at the designated time. Temporal accuracy is not usually a concern for machine tools. However, for advanced LPBF scan strategies incorporating line-to-line or within-line velocity or power control, both temporal and spatial accuracy are essential. To accomplish this, jerk-limited motion control is implemented on the NIST AMMT.

2.1. Jerk-limited path design

A sine function is chosen for the jerk. Jerk is the time derivative of acceleration, therefore can be integrated with boundary conditions to get the path (position) profile, *x(t)*, where *t* is the time. The detail equations are given below:

$$j(t) = K \cdot sin(\omega t) \tag{1}$$

j(t) is the jerk, K is its amplitude, ω is its angular velocity

$$a(t) = \int j(t)\, dt = -\frac{K}{\omega}\cos(\omega t) + C \tag{2}$$

a(t) is the acceleration, C is a constant; at $t = 0$, $a(t) = 0$, $C = K/\omega$.

$$v(t) = \int a(t)\, dt = -\frac{K}{\omega^2}\sin(\omega t) + \frac{Kt}{\omega} + D \tag{3}$$

v(t) is the velocity, *D* is a constant; at $t = 0$, $v(t) = 0$, $D = 0$.

$$x(t) = \int v(t)\, dt = \frac{K}{\omega^3}\cos(\omega t) + \frac{Kt^2}{2\omega} + E \tag{4}$$

x(t) is the position, *E* is a constant; at $t = 0$, $x(t) = 0$, $E = -K/\omega^3$. Setting a constraint of maximum acceleration, *A*, allowed on the system using Eq. 2 yields:

$$\frac{2K}{\omega} = A \tag{5}$$

At $t = 2\pi/\omega$, set $v(t) = F$, where F is the feed rate. From Eq. 3

$$\frac{2K\pi}{\omega^2} = F \tag{6}$$

Solving Eq. 5 – 6 for ω and K, the path equation will be

$$x(t) = \frac{F^2}{2A\pi}\, cos\left(\frac{A\pi t}{F}\right) + \frac{At^2}{4} - \frac{F^2}{2A\pi} \tag{7}$$

2.2. Path planning comparison

Nine square laser scan paths (each consisting of four sequential moves along the sides of a 4 mm by 4 mm square) were generated on the AMMT using different motion control parameters (Table 1). Laser power is a constant 100 W. For the step velocity profile, a wait time was introduced after each move to improve spatial path accuracy, emulating commercial controllers.

Yeung, Ho; Lane, Brandon; Donmez, M; Fox, Jason; Neira, Jorge.
"Implementation of Advanced Laser Control Strategies for Powder Bed Fusion Systems."
Paper presented at 46th SME North American Manufacturing Research Conference, NAMRC 46, College Station, TX, United States. June 18, 2018 - June 22, 2018.

Table 1. Parameter settings for squares scanned with different motion controls. The step velocity is simulated by a 100 000 m/s2 acceleration.

Scan #	Velocity Profile	Max. Acc.	Wait time
1	Sine vel. (Jerk-limited)	1000 m/s^2	0 s
2	Sine vel. (Jerk-limited)	5000 m/s^2	0 s
3	Sine vel. (Jerk-limited)	100 000 m/s^2	0 s
4	Ramp vel.	1000 m/s^2	0 s
5	Ramp vel.	5000 m/s^2	0 s
6	Ramp vel.	100 000 m/s^2	0 s
7	Step vel. with wait time	100 000 m/s^2	0.002 s
8	Step vel. with wait time	100 000 m/s^2	0.005 s
9	Step vel. with wait time	100 000 m/s^2	0.0005 s

The image of the scan tracks is shown in Fig. 1a, and the commanded and measured scan paths are plotted in Fig. 1b, with scan numbers marked on the figures. The distortion of the scanned squares occurs when the next move starts before the current destination can be reached. A carefully calibrated wait time can be introduced to compensate this distortion, such as shown in scan 7. However, this wait time improves only geometric accuracy and it is velocity sensitive (Sec. 4). If it is too long, it will cause over melting (Fig. 1a scan 8, red arrows). If it is too short, it cannot fully compensate the distortion (scan 9).

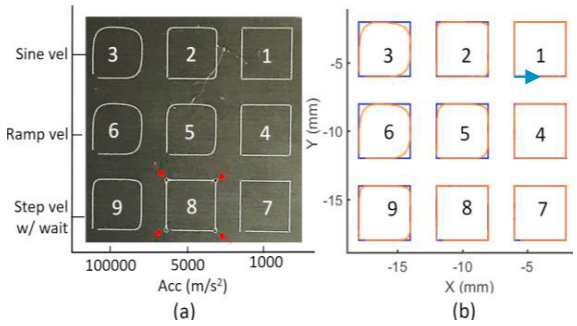


Fig. 1. Square scan paths generated with different motion controls. (a) Image of the scan tracks on an aluminum plate. Note the acceleration scale does not apply to step velocity. (b) X-Y plot of the scan paths. Blue is the command; orange is the measured. The scan is in counter-clockwise direction. The blue arrow marks the first side scanned.

The commanded and measured x-axis position $x(t)$, velocity $v(t)$, and acceleration $a(t)$ for scans 1, 4, and 7 in Table 1 are plotted in Fig. 2. Step velocity requires an impulse acceleration, which is unrealistic on any physical system, and shows the greatest deviation from the commanded path. Ramp velocity is much better followed except at the corners. Sine velocity (jerk-limited) is best followed. Wait time can be added in all cases to improve spatial accuracy, but has no effect on temporal accuracy.

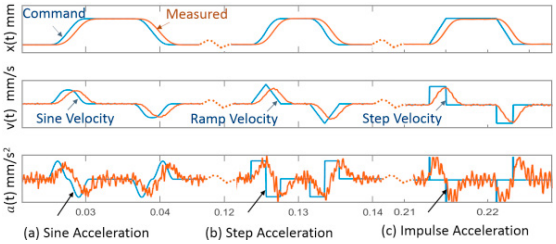


Fig. 2. Position $x(t)$, velocity ($v(t) = dx(t)/dt$), and acceleration ($a(t) = dv(t)/dt$) plotted against time for galvo x axis. Blue is the command; orange is the measured.

2.3. Temporal accuracy of scan path

To visualize the effect of temporal path accuracy, two series of 2 mm x 2 mm patterns were scanned on an aluminum plate at different speeds (200 mm/s to 2000 mm/s) with jerk-limited and step velocity motion controls. Constant build speed, constant power modes (section 4.1) were used; hence the laser power turns on and off at designated positions. Acceleration for jerk-limited control was set to 1000 m/s^2. Wait time for step velocity control was calibrated at speed of 200 mm/s. Figure 3 shows the scan tracks on the aluminum plate. The gaps in the scanned patterns for step velocity control indicate the laser spot did not reach the designated position at the designated time (i.e., a temporal error). No gap was observed for jerk-limited control at all speeds.

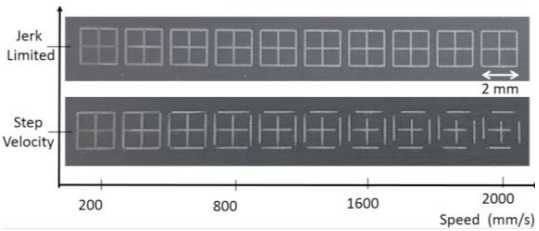


Fig. 3. Comparison of jerk-limited and step velocity motion control at different speeds.

3. Influence of laser control on LPBF process

Laser power and scan speed influence the input energy density, and any errors resulting from control lead to non-uniformities in process characteristics, potentially causing material defects. Since one key signature of process characteristics is the geometry (size and shape) of the melt-pool, one can observe the

Yeung, Ho; Lane, Brandon; Donmez, M; Fox, Jason; Neira, Jorge.
"Implementation of Advanced Laser Control Strategies for Powder Bed Fusion Systems."
Paper presented at 46th SME North American Manufacturing Research Conference, NAMRC 46, College Station, TX, United States. June 18, 2018 - June 22, 2018.

melt-pool to study the influence of laser control on the process. To measure melt-pool geometry, a high-speed camera was setup coaxially with the laser beam using a dichroic mirror, imaging lens, and filter. Emitted light from the melt-pool, which is filtered at 850 nm (40 nm bandwidth), is imaged on the camera sensor with nominal 1:1 magnification and 12 µm pixel size. The camera is set to 30 000 frames/s, 31.6 µs exposure time, 256 pixel x 256 pixel window, and 8-bit dynamic range (grayscale). The gray levels are used to relate to melt-pool dimensions [13]. Contours, representing isotherm lines, can be drawn on the raw melt-pool image to represent equal intensity (Fig. 4). A contour with intensity digital level of 170 was found to equate to the physical melt pool width based on the ex-situ measured scan track width via microscope inspection. This digital level contour is then used to infer melt pool boundary from the high-speed images and calculate melt pool dimensions and area.

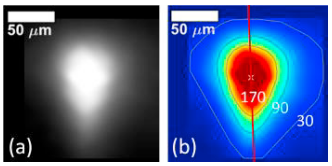


Fig. 4. Melt-pool image analysis. (a) raw grayscale image. (b) processed image. Black contour lines show different intensity levels (DL); red line shows melt-pool orientation.

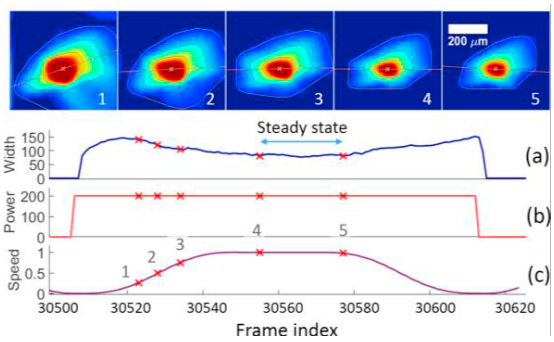


Fig. 5. A single-track scan on stainless steel. (a) Melt-pool width (µm) measured from in-situ melt-pool images. (b) Commanded laser power (W). (c) Commanded laser speed (m/s). Melt-pool images corresponding to the marked locations (1-5) are shown on the top.

An example of using in-situ melt-pool imaging to study the effects of laser control on melt-pool geometry is shown in Fig. 5. A single track was scanned on stainless steel and monitored using the co-axial high-speed camera. The melt-pool width measured from images is plotted together with commanded laser power and laser speed (Fig. 5a – 5c). The images at speed = 0.25 m/s, 0.5 m/s, 0.75 m/s, and 1 m/s are shown on top of the plots, with their respective locations (1-5) marked. It can be seen from Fig. 5a that melt-pool width decreases as speed increases (1-4), and is relatively constant at constant speed (4-5). For a uniform process, the melt-pool size must be kept constant, which can be achieved by changing the laser power in coordination with the instantaneous velocity and position of the scan. For such coordination, high temporal accuracy of scan velocity is required.

4. LPBF scan strategies and implementation by G-code

The jerk-limited path planning makes it possible to develop complicated scan strategies, which require precise time-velocity and time-position relationship. Such strategies include modulating laser power with instantaneous velocity to achieve constant power density, or modulating power with instantaneous location to respond to dynamic thermal effects stemming from heat accumulation due to local variations in part geometry or scan history. The melt-pool continuity may also be important. The on/off modulation of the laser and dramatic variation of laser power or speed can perturb a nominally steady-state melt-pool. A more 'smooth' build may be possible if there are reduced power and speed variations. To facilitate the test and implementation of complicated scan strategies, we proposed the concept of laser path modes and laser power modes [10], and implemented them through a modified version of G-code (referred as AM G-code).

4.1. AM G-code

G-code is a high-level programming language for computer numerical control (CNC). An EIA standard for G-code can be found in [14]. A simple G-code line such as 'G01 X1 Y1 F1000' commands the machine tool to move linearly (G01) from current position to x-y coordinate (1, 1), with steady state feed (F) of 1000 (mm/s). Such a 'move' is interpolated by G-code interpreter into a sequence of digital positions (micro-steps sent to motor controller) based on the velocity profile and the path mode. The path mode defines how sequential moves are planned. For example, for a rectangular path such as in Fig. 1, the current move can stop completely before next move; or it can continue

Yeung, Ho; Lane, Brandon; Donmez, M; Fox, Jason; Neira, Jorge.
"Implementation of Advanced Laser Control Strategies for Powder Bed Fusion Systems."
Paper presented at 46th SME North American Manufacturing Research Conference, NAMRC 46, College Station, TX, United States. June 18, 2018 - June 22, 2018.

to next move through a connection arc. Hence for same G-code, different paths can be interpreted based on different path modes.

Conventional G-code (and M-code) does not support power control within a move. The AM G-code is developed by adding a keyword 'L' to specify the laser power level. The usage of L is similar to the keyword F (feed). Power mode is also defined and together with L to describe the laser power profile within a move. A summary of laser path and power modes defined for AM G-code is listed below.

4.2. AM G-code interpretation modes

Three laser path modes and three laser power modes are defined. A combination of both can be used to fully describe the power-velocity-position strategy.

Laser path modes
1) Exact stop – complete stop at the end of each move.
2) Constant build speed – keep speed constant while laser is on.
3) Continuous – match the end and start velocity of two moves.

Laser power modes
1) Constant power – keep laser power constant during each move.
2) Constant power density – keep power/speed ratio constant.
3) Thermal adjusted – adjust power per predefined/ determined thermal properties or feedback from real-time monitoring.

One laser path mode and one or multiple (such as constant power density + thermal adjusted) laser power modes can be set for interpretation - hence the same AM G-code script can be interpreted into different scan strategies. As an illustration, a matrix of nine 2 mm x 1.5 mm rectangular areas was interpreted with the 3x3 combinations of the laser path and power modes (Fig. 6-7). The areas are filled by a hatching pattern (rastering) of 0.2 mm spacing and 45° inclination angle. The color bars in the figures indicate laser speed and power, respectively. The scan sequence is numbered 1–9. The constant build speed mode is implemented by allowing the overshoot of the path but with laser power turned off; the continuous mode replaces the sharp corner with an arc; the constant power density mode maintains constant power-to-velocity ratio; and the thermal adjusted mode by inversely proportion power to 'proximity'. 'Proximity' definition is based on the distance from the neighborhood points already scanned. Similarly, thermal adjusted mode can also be implemented based on local geometry and heat conduction (section 4.3). The usage of laser power keyword L is further illustrated in path 6 (Fig. 7), where L200 (laser power = 200 W) was set for all linear moves and L100 was set for the connection arcs.

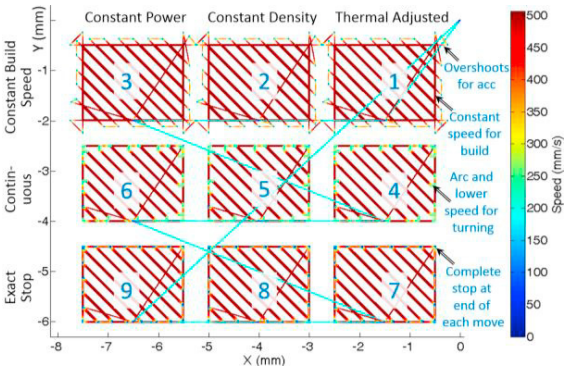


Fig. 6. Path planned by the combinations of three laser path and three laser power modes. Laser speed is represented by color.

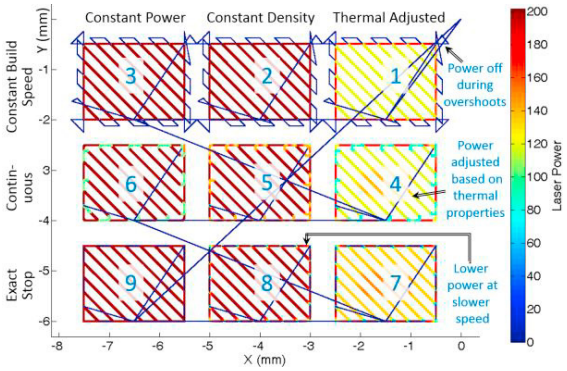


Fig. 7. Path planned by the combinations of three laser path and three laser power modes. Laser power is represented by color.

4.3. Implementation of thermal adjusted mode

The thermal adjusted mode demonstrated in Fig. 6-7 is based on a single layer residual heat compensation model. The similar concept can be extended to more complicated multiple-layer builds such as the overhanging structure (a bridge) shown in Fig. 8a. Overhanging structure is problematic to build because the large variation in thermal conductivities between powder and solidified regions. Traditionally this is addressed by either adding support structures to improve the local thermal conductivity [15], or changing the structure design itself [16]. The thermal adjusted mode proposed here provides a framework to handle such issues through fine tuning of laser power

Yeung, Ho; Lane, Brandon; Donmez, M; Fox, Jason; Neira, Jorge.
"Implementation of Advanced Laser Control Strategies for Powder Bed Fusion Systems."
Paper presented at 46th SME North American Manufacturing Research Conference, NAMRC 46, College Station, TX, United States. June 18, 2018 - June 22, 2018.

and velocity at each scan point. Since it operates at G-code interpretation level, it is independent of the structure design, and hence a more generic solution. A unitless geometric-based thermal conduction factor (GCF) is developed in the interpreter which is conceptually demonstrated in Fig. 8.

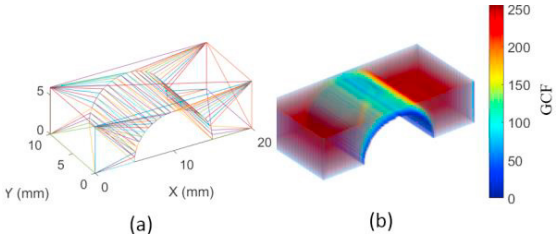


Fig. 8. Thermal adjusted mode implementation. (a) STL plot of a bridge structure. (b) Geometric conductivity factor (GCF) model constructed from the scan path.

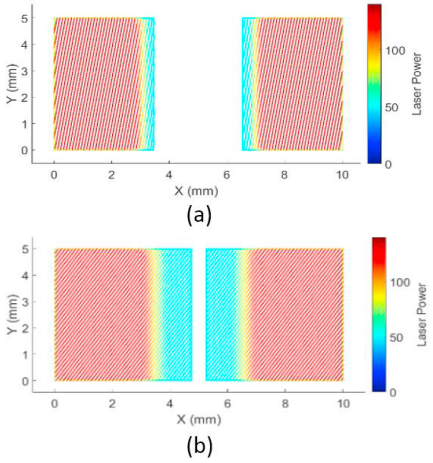


Fig. 9. Scan power at (a) 200th layer. (b) 250th layer. The laser power is reduced gradually at the overhanging area.

The x-y scan positions generated by the G-code interpreter for the part in Fig. 8a are used to create a layer-wise bitmap for the 'melted' pixels. A pixel is 'melted' if it is within a specified distance of the laser spot center at a 'laser on' scan point. Depending on the pixel size defined (10 µm by 10 µm pixel is used here), a relatively precise cross section of the part being built can be modeled. These bitmap layers are then added up layer by layer, and a GCF value is assigned to the current pixel based on the weighted GCF value of the already-built pixels (from previous layer and its same layer neighbors) with immediate contact to it. Pixels on the base plate (0^{th} layer) have a full GCF value. A weighing factor is based on a hypothetical cylinder with diameter approximately equal to melt pool width, and depth equal to powder layer. For example, a 100 µm melt-pool and 25 µm layer results in a 50 % weight to the previous layer since ratio of the bottom surface area to side surface area is about 50 : 50.

A multi-layer GCF model (or a three-dimensional GCF lookup table) can hence be built. Fig. 8b shows such a model for the object in Fig. 8a. Once this model is built, the laser power at each scan point can be adjusted according to the GCF value at that location. A linear function $L = L_o (aX+b)$ can be used to adjust the laser power, where L is the adjusted laser power, L_o is the original laser power, X is the normalized GCF, a and b are constants which can be optimized from experiments. Figure 9 shows the adjusted laser power at different layers for the bridge structure in Fig. 8. Note the gradually decreasing power level when approaching the overhanging region.

5. Comparison of different scan strategies

A key signature characteristic in LPBF AM processes is the melt-pool geometry. It is used to compare the effects of different scan strategies in this study. In-situ high-speed coaxial imaging is used to measure the melt-pool image area, and ex-situ confocal microscopy is used to measure the surface topology of the solidified melt-pool (scan track).

5.1. Melt-pool image area

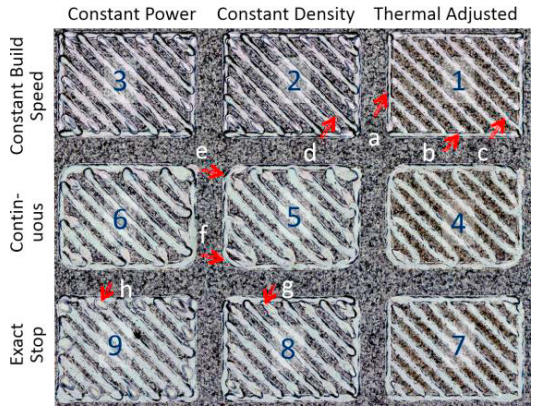


Fig. 10. Path planned by the combinations of three laser path and three laser power modes. Image shows the scan on a stainless-steel plate.

The paths planned in Fig. 6-7 were scanned on a stainless-steel plate. The images of the scanned areas in Fig. 10 provide an overview of the effect of different scan modes. The result from in-situ melt-pool size

analysis is given in Fig. 11, with camera images demonstrated from locations 'a' to 'h' marked on both Fig. 10 and 11. The melt-pool image area is plotted together with laser power and scan speed. For constant build speed mode (1-3), the laser was turned on and off at very precise locations. No laser marks outside the rectangles are visible, and no gap at the connection of the scan tracks are observed. The power-velocity-position is well coordinated under jerk limited motion control. Constant build speed mode has the most even melt-pool size. Continuous mode (4-6) creates a continuous melt-pool. But, the irregularity in the melt-pool image area will require further effort to improve. Exact stop constant power mode (9) shows the biggest variation of melt-pool size. However, the melt-pool is much more consistent in the exact stop constant density mode (8) and exact stop thermal adjusted mode (7), proving the power adjustment can effectively suppress the melt-pool irregularity. The thermal adjusted mode (1, 4, 7) has demonstrated the benefit of implementing more advanced scan strategies. The overshoots in melt-pool image area are reduced (e.g., comparing 7 to 8 and 9). The track width for 1, 4, and 7 are more uniform. A similar concept can be applied to overhanging structures, thin wall structures, etc. in 3D geometry to compensate for the varying local thermal conduction.

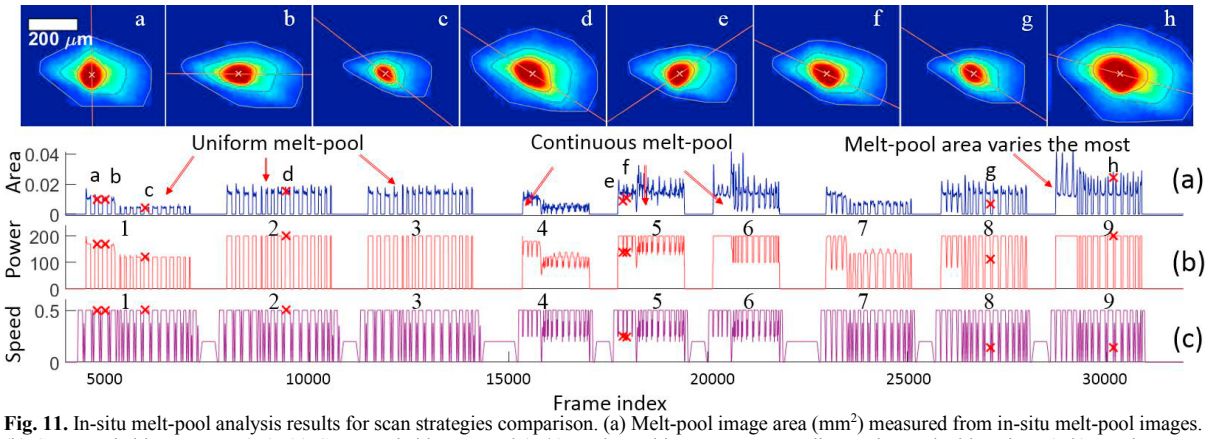


Fig. 11. In-situ melt-pool analysis results for scan strategies comparison. (a) Melt-pool image area (mm^2) measured from in-situ melt-pool images. (b) Commanded laser power (W). (c) Commanded laser speed (m/s). Melt-pool images corresponding to the marked locations (a-h) are shown on the top.

5.2. Scan track surface topology

Confocal microscopy enables the reconstruction of three-dimensional surfaces from a set of images obtained at different focal depths. It was used to compare the surface topology of the scan tracks resulted from various laser path modes. Figure 12 shows images of three single tracks scanned on a metal plate with exact stop, constant build speed, and continuous path modes, respectively. The laser power was constant mode at 200 W, and the nominal laser speed was 500 mm/s. The corresponding confocal microscopic measurements are plotted in Fig. 13. Figure 13a is the height profiles measured along the scan tracks, Fig. 13b is the surface topology of the scan track at various locations marked by the red dotted lines in Fig. 13a. A bump and a hole are clearly visible for exact stop and constant build speed modes, at the positions when the laser power was turned on / off. This agrees very well with the melt flow simulation in [6]. The hole is deeper for constant build speed mode, likely due to the fact that the laser was still travelling at high speed when it was turned off. However, this does not seem to have the same effect when the laser was turned on - the heights of the bumps for constant build speed and exact stop modes are similar. There are very little variations in the scan track heights for the continuous mode. Comparing Fig. 13 and Fig. 11, it is interesting to see constant build speed mode has the most uniform melt-pool image area but the largest variation in scan track height at the end points. On the other hand, continuous mode has the largest variation in melt-pool image area but most uniform scan track height.

The variation of track height is mainly due to the laser power switching on and off, which is the most frequent and drastic (while the laser is travelling at full speed) in constant build speed mode. The variation of

Yeung, Ho; Lane, Brandon; Donmez, M; Fox, Jason; Neira, Jorge.
"Implementation of Advanced Laser Control Strategies for Powder Bed Fusion Systems."
Paper presented at 46th SME North American Manufacturing Research Conference, NAMRC 46, College Station, TX, United States. June 18, 2018 - June 22, 2018.

melt-pool image area is mainly due to the changes of laser speed, power and direction. Constant build speed mode turns off laser power when direction changes or speed slows to minimize melt-pool area variation. Continuous mode, on the other hand, keeps laser power on all the time to minimize track height variation. If the goal is to keep both the height and area variations minimum, a combination of continuous path mode and constant power density mode seem the best choice. However, the melt-pool dynamic is very complicated, constant power density alone cannot guarantee a constant melt-pool area. Many other factors, such as thermal properties, powder dimension, gas flow, etc. can all affect the build quality. Ongoing studies are needed to continuously optimize the control parameters based on the framework proposed here.

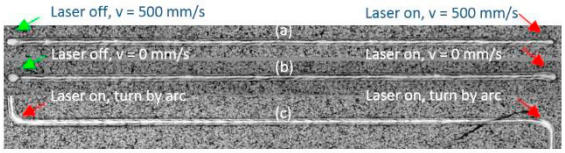


Fig. 12. Single track scan with (a) exact stop, (b) constant build speed, and (c) continuous laser path mode.

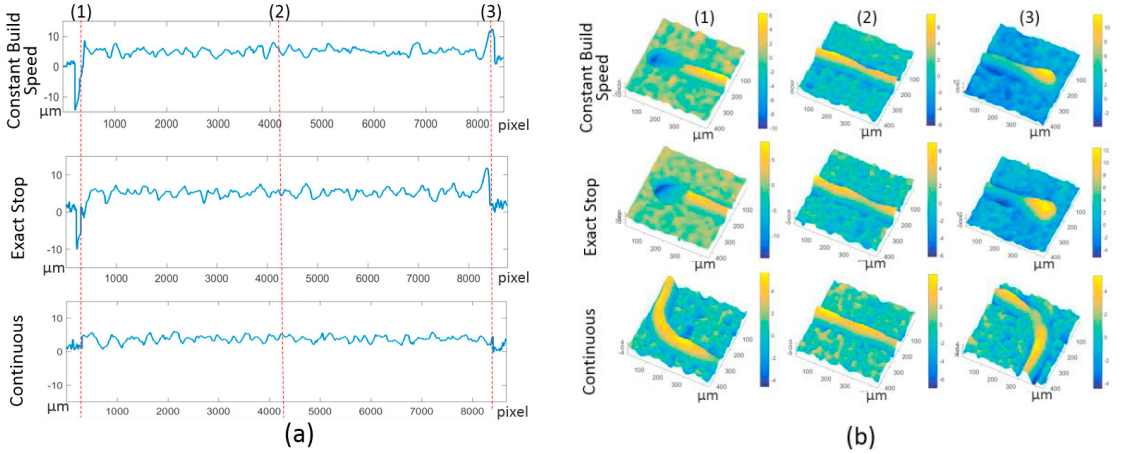


Figure 13. Confocal microscopic measurements. (a) Height profile along the melt-track. (b) The surface topographies at the locations indicated by red dotted lines in (a).

6. Discussion and summary

A jerk-limited motion control was implemented on a LPBF AM testbed, and improved position and velocity temporal accuracies were demonstrated. This enabled the implementation of advanced laser control strategies based on precise power-velocity-position coordination. Such strategies were proposed and implemented through 'AM G-code' with three laser path modes and three laser power modes built into its interpreter. A thermal-adjusted mode was also proposed that locally varies power based on adjacent solidified material and variation in local heat conduction. Scan experiments were conducted on a metal plate to demonstrate the effectiveness of different modes, in-situ melt-pool imaging and ex-situ confocal microscopy were utilized to study the processes. The melt-pool controllability is clearly demonstrated. Experiments will be conducted for multilayer powder 3D builds to further verify their effects on the quality of the built parts. Further study is still needed to understand optimal control strategies pertaining to the AM fabrication process; however here we demonstrated methods for controllability.

References

[1] Kruth J-P, Leu M C and Nakagawa T 1998 Progress in Additive Manufacturing and Rapid Prototyping *CIRP Ann. - Manuf. Technol.* **47** 525–40

[2] Mani M, Lane B, Donmez A, Feng S, Moylan S and Fesperman R 2015 *Measurement Science Needs for Real-time Control of Additive Manufacturing Powder Bed Fusion Processes* (National Institute of Standards and Technology)

[3] King W E, Barth H D, Castillo V M, Gallegos G F, Gibbs J W, Hahn D E, Kamath C and Rubenchik A M 2014 Observation of keyhole-mode laser melting in laser powder-bed fusion additive manufacturing *J. Mater. Process. Technol.* **214** 2915–25

Yeung, Ho; Lane, Brandon; Donmez, M; Fox, Jason; Neira, Jorge.
"Implementation of Advanced Laser Control Strategies for Powder Bed Fusion Systems."
Paper presented at 46th SME North American Manufacturing Research Conference, NAMRC 46, College Station, TX, United States. June 18, 2018 - June 22, 2018.

[4] Thijs L, Verhaeghe F, Craeghs T, Humbeeck J V and Kruth J-P 2010 A study of the microstructural evolution during selective laser melting of Ti–6Al–4V *Acta Mater.* **58** 3303–12
[5] Yadroitsev I, Thivillon L, Bertrand P and Smurov I 2007 Strategy of manufacturing components with designed internal structure by selective laser melting of metallic powder *Appl. Surf. Sci.* **254** 980–3
[6] Khairallah S A, Anderson A T, Rubenchik A and King W E 2016 Laser powder-bed fusion additive manufacturing: Physics of complex melt flow and formation mechanisms of pores, spatter, and denudation zones *Acta Mater.* **108** 36–45
[7] Zaeh M F and Ott M 2011 Investigations on heat regulation of additive manufacturing processes for metal structures *CIRP Ann. - Manuf. Technol.* **60** 259–62
[8] Antony K, Arivazhagan N and Senthilkumaran K 2014 Numerical and experimental investigations on laser melting of stainless steel 316L metal powders *J. Manuf. Process.* **16** 345–55
[9] Lane et al. Design, Developments, and Results from the NIST Additive Manufacturing Metrology Testbed (AMMT) *Solid Free. Fabr. 2016 Proc. 26th Annu. Int. Solid Free. Fabr. Symp*.
[10] Yeung H, Neira J, Lane B, Fox J and Lopez F Laser Path Planning and Power Control Strategies for Powder Bed Fusion Systems *Solid Free. Fabr. 2016 Proc. 27th Annu. Int. Solid Free. Fabr. Symp.*
[11] Erkorkmaz K and Altintas Y 2001 High speed CNC system design. Part I: jerk limited trajectory generation and quintic spline interpolation *Int. J. Mach. Tools Manuf.* **41** 1323–45
[12] Zhang Q, Li S and Guo J 2012 Smooth time-optimal tool trajectory generation for CNC manufacturing systems *J. Manuf. Syst.* **31** 280–7
[13] Rombouts M, Kruth J-P, Froyen L and Mercelis P 2006 Fundamentals of selective laser melting of alloyed steel powders *CIRP Ann.-Manuf. Technol.* **55** 187–92
[14] Electronic Industries Association Interchangeable Variable Block Data Format for Positioning, Contouring, and Contouring/Positioning Numerically Controlled Machines *EIA Stand. EIA-274- Febr. 1979*
[15] Järvinen J-P, Matilainen V, Li X, Piili H, Salminen A, Mäkelä I and Nyrhilä O 2014 Characterization of Effect of Support Structures in Laser Additive Manufacturing of Stainless Steel *Phys. Procedia* **56** 72–81
[16] Atzeni E and Salmi A 2015 Study on unsupported overhangs of AlSi10Mg parts processed by Direct Metal Laser Sintering (DMLS) *J. Manuf. Process.* **20** 500–6

Yeung, Ho; Lane, Brandon; Donmez, M; Fox, Jason; Neira, Jorge.
"Implementation of Advanced Laser Control Strategies for Powder Bed Fusion Systems."
Paper presented at 46th SME North American Manufacturing Research Conference, NAMRC 46, College Station, TX, United States. June 18, 2018 - June 22, 2018.

Proceedings of the ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2017
August 6-9, 2017, Cleveland, Ohio, USA

DETC2017-67807

A DOMAIN-DRIVEN APPROACH TO METAMODELING IN ADDITIVE MANUFACTURING

Zhuo Yang, Thomas Hagedorn, Douglas Eddy, Sundar Krishnamurty, Ian Grosse
University of Massachusetts Amherst
Department of Mechanical and Industrial Engineering
Amherst, MA 01003
Email: zhuoyang@engin.umass.edu, thagedorn@engin.umass.edu, dceddy@engin.umass.edu, [skrishna, grosse]@ecs.umass.edu

Peter Denno, Yan Lu, Paul Witherell
National Institute of Standards and Technology
Engineering Laboratory
Gaithersburg, MD 20899
Email: [peter.denno, yan.lu, paul.witherell]@nist.gov

ABSTRACT

Recent studies have shown advantages to utilizing metamodeling techniques to mimic, analyze, and optimize system input-output relationships in Additive Manufacturing (AM). This paper addresses a key challenge in applying such metamodeling methods, namely the selection of the most appropriate metamodel. This challenge is addressed with domain-specific AM information, derived from physics, heuristics and prior knowledge of the process. Domain-specific input/output models and their interrelationships are studied as a basis for a domain-driven metamodeling approach in AM. A metamodel selection process is introduced that evaluates global and local modeling performances, with different AM datasets, for three types of surrogate metamodels (polynomial regression (PR), Kriging, and artificial neural network (ANN)). A salient feature of this approach is its ability to seamlessly integrate domain-specific information in the model selection process. The approach is demonstrated with the aid of a metal powder bed fusion (PBF) case study and the results are discussed.

1. INTRODUCTION

AM techniques have promising applications in such different domains as aerospace, medical devices, and heavy industry [1]. However, these often multi-physical processes are still not fully understood or controlled [2]. For example, in the powder bed fusion (PBF) process, the relative density of AM parts is affected by various user controllable parameters such as laser power, scan speed, or powder density[3]. In some instances, the hardness of AM parts might be different despite being produced by same machine using the same input parameters [4]. The difficult-to-predict performances of AM processes introduce significant uncertainty into quality control and engineering design [5]. To mitigate this uncertainty, researchers are seeking mathematical predictive models that can predict material properties prior to production by the manufacturing process. Pure physics-based models [Pal et al, 2013], numerical simulation models [6], and metamodels [7] have been proposed to predict AM behavior. This paper discusses a novel method that leverages domain knowledge to improve the selection of metamodels for AM predictive modeling problems.

Metamodels, also known as surrogate models, construct a model of a model to understand complex systems [8]. Unlike physics-based or numerical simulation models that often require detailed knowledge of internal processes such as problem physics [5], metamodeling techniques focus on the input/output relationships[9]. Metamodels can significantly reduce the cost of organizing knowledge for a poorly understood system. However, these models can introduce modeling uncertainties due to lack of representative process knowledge [10], a hindrance our domain-driven technique seeks to address. Indeed, some problem-specific knowledge is desired to select an appropriate

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"A DOMAIN-DRIVEN APPROACH TO METAMODELING IN ADDITIVE MANUFACTURING."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2017, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

sampling strategy and account for system behaviors and sensitivity properties [11, 12], which can fundamentally affect model performance.

The term model performance in this paper refers to predictability and efficiency that is believed to be correlated to data sets such as sample size, data gradient, and kurtosis [13]. Previous modeling selection frameworks have used information about data, called data features, to identify the optimal modeling algorithm for certain types of problem. "Data features" in this context are considered characterizations of how each parameter impacts the responses in a given domain. For example, Rice et al. proposed a model can use extracted characteristics and performance measurements such as normalized root mean square error (NRMSE) [13] and maximum relative error magnitude (MREM) [14] from a given dataset to select the optimal modeling algorithm from a set of candidates [15]. Cui et al.'s energy model recommendation framework uses dual performance evaluation criteria and criteria reduction methods to implement a meta-learning procedure for modeling algorithm selection [13]. While these well-established methods work for some general problems or problems in specific domains, they usually require significant work in data feature characterization.

This paper aims to address AM predictive modeling challenges by constructing a specific-to-AM framework that looks to efficiently and accurately identify optimal metamodeling methods for given problems, prior to deeply investigating data features related to a specific domain. The term "domain" in this paper indicates the topic area to which the parameters apply. For example, laser power is a parameter in the AM thermal domain and powder density resides in the AM material domain, etc. In the following sections, we will first discuss the performance of metamodels with different domain-inspired AM input/output parameters. Unlike Rice's model that uses data features for algorithm recommendation, the method proposed in this paper focuses on using correlations between AM parameters to efficiently identify data features. An AM input/output correlation chart was developed to visually present the nonlinearity of different combinations of parameters. We introduce the domain-driven framework in Section 5. For demonstrative purpose, a case study based on AM datasets is presented in Section 6. The benefits of using this domain-driven approach are explained and further discussed in these sections.

Section 2 provides a brief overview of predictive models in AM. Section 3 discusses parametric correlations in AM models based on a detailed literature review. The remaining sections are built upon the findings from Section 3. Section 4 provides a brief background of metamodeling algorithms. The three candidates presented in this paper are used to demonstrate the framework. The fundamental structure of the framework and an illustrative case study with empirical data from an AM process are introduced in Sections 5 and 6. Section 7 concludes with a comprehensive discussion of the early framework.

2. OVERVIEW OF AM PREDICTIVE MODELING

Porosity, relative density, surface roughness, geometric accuracy, hardness, and tensile strength are examples of critical properties that define AM product quality [1]. While important to physics based models, values of intermediate parameters such as melt pool width, penetration depth, and melting temperature cannot represent critical AM properties directly [16]. Advancement in AM processes and product design quality requires a clear understanding of the relationship between various AM input and output parameters.

Several studies have explored the effect of AM process parameters on part performance [17]. Witherell et al (2014) analyzed the metal PBF process and divided it into four critical categories that summarize the complex inter-relationships between AM parameters [7]. In this approach, the PBF process is modeled as a set of sub-processes, including heat source/absorption model, a melt pool formation model and a solidification model, where each physical process also has multiple sub-processes. For example, Marangoni, capillary, and heat convection/conduction models are sub-sets of a larger melt pool model [18].

Other studies focus on the details of sub-processes such as heat source/absorption and melt pool models. For instance, the isotherm migration method uses theoretical analysis to estimate the surface temperature of the powder bed, which can then be used to study phase changing during 3D printing [19]. Finite element analysis (FEA) can also simulate melt pool information [20, 21]. Ma et al (2015)'s 3D FEA model of melt pool for example used user controllable inputs (laser power, scan speed, powder density, etc.) to predict melt pool width and depth [22]. These studies contribute to improved AM process knowledge, but they do not completely characterize the full set of physical phenomena in an AM processes. As a result, there is interest in using robust and efficient design of experiments (DOEs) and metamodeling approaches to study AM problems.

Metamodeling methods use a statistical approach that treats a complex system as a black box to avoid limitations stemming from lack of knowledge during model construction [8]. The data that is used to construct metamodels can be collected from computer simulations or actual measurements from DOE [23]. In general, one can improve the metamodel by adapting sampling strategies that are well suited to a specific problem [Shao et al, 2008], optimizing modeling parameters [24], and selecting a more appropriate modeling method [15].

Optimal sampling can potentially inform a metamodel with information about the unknown system [25]. For example, Shao and Krishnamurty's model updating method selects sample points that are closer to local optima to improve the predictive ability of a Kriging model [26]. Selection of modeling parameters can also affect the model performance within the same given sample set [27] [Yang et al, 2017]. Both methods assume that a pre-selected metamodeling algorithm is appropriate, which is not necessarily true. For example, it can be expected that a linear model would have difficulty producing accurate predictions for highly nonlinear problems regardless of sampling technique or model parameters. Model selection usually requires significant work to characterize potentially significant data features in a given dataset [13] or uses a complete, multi-stage adaptive sampling process [26].

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"A DOMAIN-DRIVEN APPROACH TO METAMODELING IN ADDITIVE MANUFACTURING."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2017, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

A rapid model selection procedure that can potentially skip investigation processes, but still provides reliable solutions, is highly desirable. In this paper, an AM metamodeling algorithm selection framework is proposed to accomplish this goal. This paper builds on our recent efforts in metamodeling for additive manufacturing, focusing on efficiency, effectiveness and optimal sampling [14]. Specifically, this work provides a basis for domain-driven enhanced metamodeling by explicating and exploiting the unknown and complicated correlations in the system behavior using a priori knowledge of the domain being studied.

3. ANALYSIS OF THE CORRELATION BETWEEN INPUT/OUTPUT OF PBF PROCESS

This paper investigates a metamodel selection method that leverages pre-existing knowledge of parametric relations instead of pure data analysis. For an identical system under the same conditions, one might perceive that the basic relation between inputs and outputs is unchanged. The following beam bending example is introduced to explain this hypothesis.

In the general case of any beam, when an analysis of beam bending is desired many of its characteristics may be unknown. For example, consider a large beam made of lattice of different orders of magnitude (Figure 1), with a load applied. Measurements of loading conditions can be taken, but not enough information can be obtained to perform a complete analysis due to the presence of too many variables.

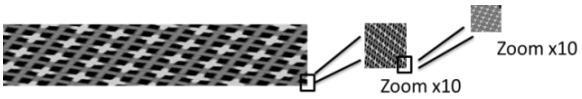


Figure 1. Beam created with lattice.

However, given the loading and constraints, we can observe that the case behaves like a bending beam. While only partial measurements can be taken at micro and meso scales, we can make the hypothesis that it will perform according to beam theory at the meso and macro levels, thus making domain-specific observations. These observations allow us to extrapolate measured values, based on a combination of measured results and expected macro performance.

A similar situation is found in metal PBF processes in that they are very complex and involve a large number of parameters (more than 50) [1, 16]. The input/output relation is difficult to discern using theoretical analysis. If one considers every single PBF parameter, there are millions of combinations of different models. Thus, section 5 will introduce a domain-driven framework that uses past model performance to predict the appropriate modeling method for a new problem. The rest of this section will summarize parametric relations between AM parameters from the literature to construct the prerequisite knowledge for development of a model selection framework.

Figure 2 shows 8 inputs and 6 outputs that have high occurrence in recent PBF literature. The relation between these parameters is currently marked as unknown due to lack of information.

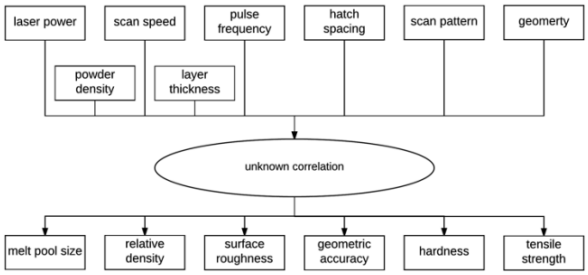


Figure 2. Infrastructure of principal AM inputs/outputs

For single input cases, PBF parameters such as laser power, scan speed, and hatch spacing have highly linear correlations to certain outputs [29, 30]. For example, Tang et al (2003)'s metal laser sintering experiment indicates the surface roughness and tensile strength is linearly increased with laser power [31]. Similarly, tensile strength decreases monotonically with higher scan speed with other parameters held constant. From the same study, however, surface roughness was not linearly related to scan speed. With reduced layer thickness, scan speed and roughness instead had a slight nonlinear relation. Another study, using different materials but the same laser parameters, found scan speed and layer thickness have a linear relation to relative density when varied individually [32]. Such similar results using multiple materials suggests that relative density is linearly related to laser power, scan speed, and layer thickness [33]. Intermediate outputs such as penetration depth (not included in Figure 2) have also been shown to exhibit a linear relation with some parameters. Kruth et al (2003)'s laser sintering experiment found higher scan speeds produced a linearly decreasing layer thickness [34]. Some inputs, such as laser pulse frequency, have been found to have a highly nonlinear relation to relative density [35].

Compared to the single input/output problem, the relations between variables become considerably more complicated when studying multiple input parameters. Tang et al (2003) found that the surface roughness is not linearly related to a combination of laser power and hatch spacing [31]. Similarly, when considering the relation of laser power and layer thickness to surface hardness, the relation is also nonlinear [32]. A similar, nonlinear relation is also observed in Morgan et al (2004)'s empirical result [35]. The simple linear relation between scan speed and relative density becomes significantly more complicated when pulse density is also varied. This evidence seems to imply that more variables generate larger uncertainties in PBF due to an increase in unknown interactions. However, other outputs such as the tensile strength are not that sensitive to the same combinations of inputs. The observed relation remains linear under a combination of factors. Thus, more inputs can increase PBF problem complexity but do not necessarily indicate increasing nonlinearity of the input/output relation.

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"A DOMAIN-DRIVEN APPROACH TO METAMODELING IN ADDITIVE MANUFACTURING."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2017, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

This paper focuses on the development of a general algorithm recommendation framework based on input/output parameters before deeply investigating physical interactions between PBF parameters. As such, it is necessary to understand the general factors that cause the uncertainty in PBF processes. Beaman et al (1997) first introduced the concept of energy density for AM, which is described by the Equation (1) [36]:

$$E_\rho = \frac{4P}{\pi r^2}\frac{2r}{v}\frac{2r}{s} \quad (1)$$

where E_ρ is Energy density, P is laser power, r is beam radius, v is scan speed, and s is hatch spacing.

Equation (1) indicates that higher laser power, lower scan speed, and closer hatch spacing produce higher energy density. More energy delivered to the powder usually means better melting conditions. Improved melting conditions will result in lower porosity and thus higher relative density. For example, Meier et al (2008)'s experiment with the metal laser sintering process shows the relative density increases from 69% to 99% with a power increase from 30W to 90W with other parameters held constant [37]. Another study [38] concluded higher energy density tends to produce a continuous melting track against irregular melt shape. These findings imply the linear relations may be more likely if the involved input/output parameters can be related to an overall energy density dependency.

In contrast, nonlinear relations were found in studies on outcomes related to part microstructure. Meier et al (2008) found surface roughness is not monotonically increased with scan speed, with the optimal roughness obtained in the middle of the range of scanning speeds tested [37]. Other research suggests that microstructure varies throughout the entire PBF part, and thus can be considered a local rather part-wide property. Wang et al (2012) found the hardness tests at different locations/directions in the same AM part produce different results [4]. Similarly, studies of thermal conditions indicate that variation of thermo-physical properties of AM parts are complicated [39].

Figure 3 summarizes the hypothetical relation of input/output correlations observed in the literature. The thick arrow in the middle of the figure represents the relation from linear to highly nonlinear. The arrow (right to left) on the top represents whether the input parameters can be classified as related to energy density or not. The bottom arrow (left to right) represents whether the outputs are in macro-scale or micro-scale. The observations performed in this section pertain to linearity of input/output relationships. Figure 4 illustrates the more general case to which these techniques may be applied.

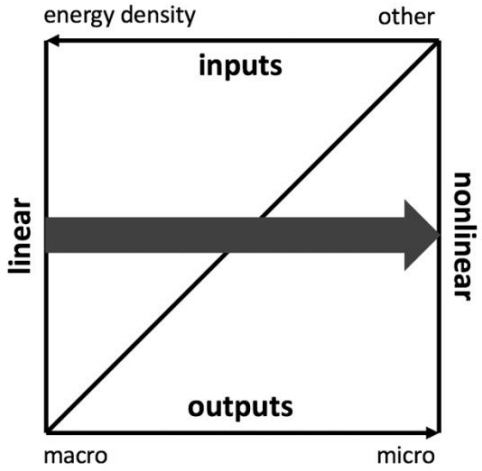


Figure 3. Hypothetical relation of input/output correlations

Figure 4.General case to model input/output relationships

Figure 3 can be used to summarize past literature results for the particular case of PBF. As discussed, laser power, layer thickness, and scan speed are input parameters that relate to energy density. Pulse frequency is located in the upper right corner since it is unrelated to energy density. For outputs, relative density is considered a high macro scale property as it depends on part width rather than local porosity. Surface roughness, however, relates more to AM microstructure. For the problem that involves the parameters in Figure 5(a), the link between inputs and outputs intersects with the bold arrow on the left. Figure 5(b) indicates pulse frequency and surface roughness have a highly nonlinear relation. Figure 5(c) and Figure 5(d) demonstrate the limitations in past research.

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"A DOMAIN-DRIVEN APPROACH TO METAMODELING IN ADDITIVE MANUFACTURING."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2017, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

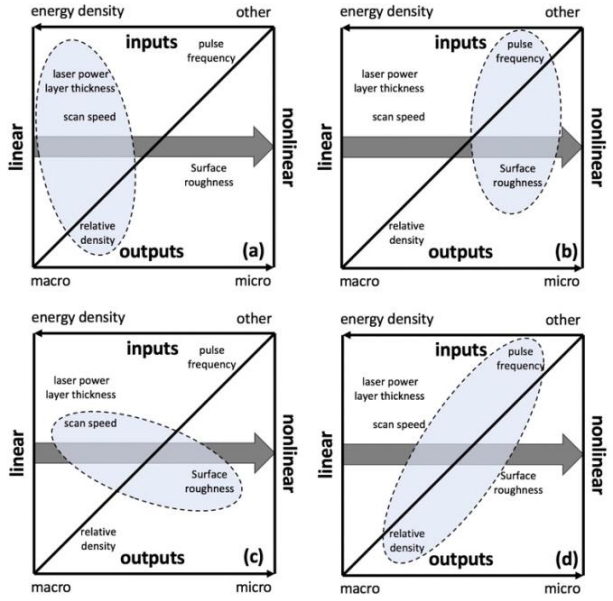


Figure 5. The use of AM input/output correlation chart

Figure 5 visually summarizes the relation between different combinations of PBF parameters observed in the literature. However, it must be stressed that, like the literature, it only summarizes some of the parameters of interest, and might not be sufficient to guide metamodel selection, especially in the case of indeterminate parameter sets. A more rigorous mathematical solution is thus needed. Section 5 introduces a framework to recommend models based on mathematical computation.

4. OVERVIEW OF METAMODELING TECHNIQUES

This section reviews the metamodeling techniques that are applied in Section 5. PR, Kriging, and ANN techniques were used in this study as they cover both parametric and non-parametric techniques [13]. However, the framework should not be limited to only these algorithms. These three were selected in this paper for demonstrative purposes. In general, any metamodeling technique can be considered as a candidate during the actual practice.

A typical metamodel [40] can be expressed as:

$$y(\tilde{x}) = f(\tilde{x}) + \varepsilon \quad (2)$$

where $y(\tilde{x})$ represents an unknown function, $f(\tilde{x})$ is a known function of $\tilde{x}$ derived statistically, and ε is the error part. $\tilde{x}$ represents the set of the system's independent input variables. For different modeling methods, the function represented by each part of the expression is different.

4.1 Polynomial Regression (PR)

The PR technique is a variation of linear regression in which an n^{th} order polynomial is used to model the relationship between the independent variables $\tilde{x}$ and the outcomes $y(\tilde{x})$ [40]. The method is popular in various engineering domains since it is fast and easy to use. A second order quadratic polynomial function has the form of:

$$\hat{y} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_i \sum_j \beta_{ij} x_i x_j \quad (3)$$

where β_0, β_i, and β_{ij} are regression coefficients, and k is the number of design variables.

4.2 Kriging

Unlike the parametric techniques producing an actual formula, the Kriging method as a non-parametric method builds its estimation based on the position of sample data. The underlying assumption of Kriging models is that an unknown point can be estimated from observed (known) points based on spatial correlation [41]. The estimation process is completed by a variogram or so called spatial correlation functions [42, 43]. The general form of a kriging estimation for an unknown predicted value of a point Z_E for a single outcome is [44]:

$$Z_E = \bar{Z} + \sum_{i=1}^{n} \lambda_i (Z_i - \bar{Z}) \quad (4)$$

where $\bar{Z}$ represents the regional mean value of the response and. λ_i is the distance-correlated weight value, which is determined by the computation of spatial correlation. The value of spatial correlation can be derived from:

$$R(\theta, x_i, x_j) = \prod_{l=1}^{n} \exp(-\theta_l (x_{i,l} - x_{j,l})^2) \quad (5)$$

where $x_{i,l}$ is the lth component of the ith vector x_i [44]. $R(\theta, x_i, x_j)$ depends on the location of points x_i and x_j, and the correlation parameter, θ. Multiple Kriging methods exist, including simple Kriging, ordinary Kriging, regression Kriging, etc. [43]. The ordinary Kriging method is used in this paper.

4.3 Artificial Neural Network (ANN)

ANN is a computational algorithm that mimics the central nervous system [45] and has been widely used for solving problems with complicated structures. A typical ANN model consists of an input layer, hidden layers, and an output layer [46]. Each layer consists of "neurons" that are connected across layers to transmit and deduce information. The optimal number of neurons and hidden layers may differ, and depends on the complexity of the problem. The structure of a simple ANN model [47] is shown in Figure 6, where x1 to x3 are input parameters, u1-4 are the neurons in the single hidden layer, and outputs y1 and y2 are located in output layer.

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"A DOMAIN-DRIVEN APPROACH TO METAMODELING IN ADDITIVE MANUFACTURING."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2017, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

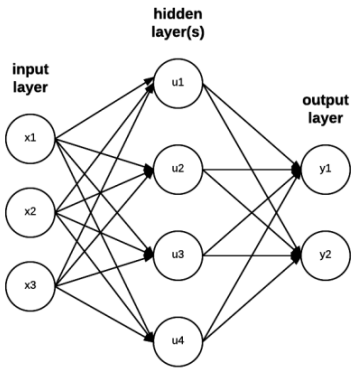


Figure 6. Typical structure of a simple ANN model

5. DOMAIN-DRIVEN MODEL RECOMMENDATION FRAMEWORK FOR AM

An exhaustive search, which is also known as the generate and test method, is the most general problem solving technique for systematically enumerating all possible candidate algorithms and selecting the most appropriate candidate based on a set of criteria [48]. While it is a global optima algorithm, it is also extremely inefficient, especially for those problems with abundant candidates and/or large input datasets. In such cases, it may be more efficient to incorporate prior knowledge into the algorithm selection process.

Many selection or recommendation techniques were developed to improve the efficiency of exhaustive search. Rice's model [15], for example, can recommend the best candidate for a new instance based on previous model selection knowledge. It includes four spaces: the problem space P represents the datasets of learning instances; domain space F contains the characteristics; algorithm space A includes all candidate algorithms; performance space Y is the measured performance of instance P for each algorithm in A [15]. Rice's model compares characteristics of a new instance to all previous examples and then assesses the suitability of each algorithm based on a set of rules or a selection algorithm. The model findings can be used to select the optimal algorithm from a given problem. Once the solution is derived, the performance in the new instance is added to the performance space Y, updating the model with a new point. In this way, a user can avoid exhaustively testing each candidate algorithm for a new instance [13].

The proposed domain-driven method is built upon Rice's algorithm selection method. However, instead of using data-features to characterize the new dataset, this proposed AM framework uses AM knowledge to indicate a possible optimal option from candidate algorithms. This approach is fundamentally different than Rice's method in that the result can be independent of the unfixed data strategy and rely on the relatively fixed knowledge of the physics of the problem. The AM characteristics used are the relations between input/output parameters discussed in section 3. The general workflow of the proposed framework is shown Figure 7. It requires sufficient knowledge to commence the selection process. Knowledge construction consists of collecting existing datasets, classifying the instances, and computing the performance of each candidate algorithm on each dataset. The extracted information is then fed back into the current knowledge model and the system predicts a possible optimal metamodeling algorithm. Before proceeding to model construction with actual data, the newly calculated solution updates the knowledge model. Details of each critical step are discussed in the rest of this section.

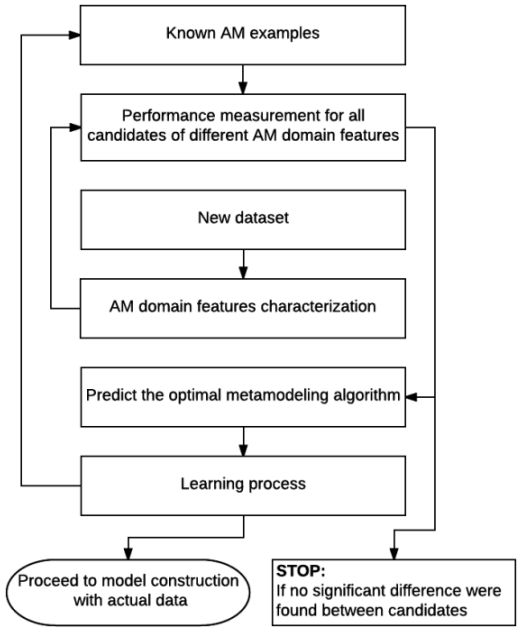


Figure 7. General workflow of the proposed framework of AM domain-driven modeling selection method

5.1 AM characterization

The AM characteristics mentioned at this stage are AM input/output parameters. At this step, all parameters are formed to input vector X of the selection framework:

$$\mathrm{X} = [x_1\ x_2\ x_3 \dots\ x_n\ y_1\ y_2\ y_3 \dots\ y_m]^T$$

where the x and y are the inputs and outputs respectively and are equal to 1 or 0. 1 indicates that the problem includes the parameter and 0 indicates the parameter is not considered. These vectors are the inputs to the learning process. For problems that have exactly the same outputs, y can be ignored.

5.2 Performance measurement

Measuring model performance of known datasets is critical to improving model selection accuracy. Two criteria were employed in the case study in section 6 to evaluate the modeling performance by a set of candidate algorithms. For global measurement, normalized root mean square error (NRMSE) [13] is used. Maximum relative error magnitude (MREM), on the

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"A DOMAIN-DRIVEN APPROACH TO METAMODELING IN ADDITIVE MANUFACTURING."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2017, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

other hand, is used to evaluate the outstanding error of the models [14]. These were formulated as:

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i-\tilde{y}_i)^2}{N}}/(y_{max}-y_{min}) \quad (6)$$

$$MREM = \max(\frac{|y_i-\tilde{y}_i|}{y_i}) \quad (7)$$

Where y_i ($\neq 0$) is the actual observed value, $\tilde{y}_i$ is the estimated value from the metamodel, y_{max}/y_{min} are the maximum/minimum actual observation, and N is the total amount of validation samples. With NRMSE and MREM, the framework can make its recommendation based on both global and local performances of the datasets by assigning appropriate weights to each criteria. In this paper, all case studies consider the NRMSE and MREM criteria equally to not bias either way. However, in some cases, these two objectives could conflict. Under these circumstances, a user could deploy a weighted multi-criteria decision making formulation.

5.3 Prediction process

The prediction process of the proposed recommendation framework could be completed by either model-based or instance-based methods. Model-based methods build predictive models to determine the optimal modeling algorithm. The predictive selection is based on an input vector X (model variables) and the resulting modeling performance (model outcome). Once the model is built, the new instance with PBF-related information in X_{new} would then import to the model and calculate the predictive result. The model-based method is similar to what is discussed in Section 3. For example, the vector X is the input variable set of the recommended predictive model. The values of NRMSE and MREM then become the predictive results. Once the recommended model is built based on existing instances, the model can predict the NRMSE and MREM of candidate algorithms for a new problem. The user can decide which algorithm would be employed according to these indicators of model performance.

An instance-based method by comparison solves the problem based on existing examples. It assumes an algorithm has similar performance on similar problems, where the similarity is measured by Euclidean distance between instance input and output vectors [Brighton et al, 2002]. The k-nearest neighbor (k-NN) ranking approach as was employed in this study. The k-NN approach ranks the nearby k nearest examples for their similarity. However, the simplest case of k-NN is the closest neighbor example based on the comparison of the Euclidean distance of all examples, which also called as 1-NN. The formulation of 1-NN is:

$$dist(i,j) = \sqrt{(a_i - a_j)^2}, \ \ j = 1, \dots, m \quad (8)$$

Where a_i represents the new instance a_j represents the existing examples, and m is the total number of examples. In the case of the metamodeling algorithm, a set of datasets composed of input and output parameters would serve as the existing points, each of which has been characterized by a set of metamodels. By comparing the input parameters (a_j) to those used in the new dataset, the user can then determine how similar the data is to a known dataset. At this point, users would simply compare the performance of various modeling algorithms in the existing examples. Thus, a likely best predictive modeling algorithm can be chosen without costly characterization of information and data features of the dataset. This saves the cost of testing all candidate algorithms individually, allowing the user to directly proceed to model construction and parameters optimization. A demonstrative example in the following section shows how the method works.

6. DEMONSTRATIVE CASE STUDY

A simple example was constructed from existing AM datasets (2 for constructing knowledge and 1 for verification) to illustrate the proposed AM domain-driven framework. Tang et al (2003)'s and Morgan et al (2004)'s metal PBF datasets were used to construct the knowledge base [31, 35]. The dataset from Chatterjee et al (2003)'s was selected to verify the selection accuracy [49]. The three datasets have the same output parameter, relative density. Because of this, the Y output vector is omitted, and the results obtained may potentially be more accurate since the knowledge is constructed from somewhat similar examples.

The 1-NN method was used to predict the optimal modeling algorithm. The predicted algorithm was then compared to actual modeling results with all candidate algorithms to assess the predictive solution [49].

6.1 Knowledge construction

The knowledge used for model predictions was composed of a small dataset with 15 samples [31] and a large dataset with 105 samples [35]. They were selected to build the knowledge model because their similarities: 1) both are metal PBF process; 2) both use similar experimental conditions; 3) they have the same output as the new dataset (relative density). The differences between them also provide opportunities for future model selection for new instances, namely: 1) they use different input parameters; 2) they use a different DOE strategy and 3) both have different variables that are not considered in DOE such as materials and specific machines. Thus, the knowledge base of these datasets is reasonable and has useful variation.

For initial construction, we consider 6 independent input variables though neither dataset can cover the parameter of layer thickness. The two datasets overlap in individual parameters. The matrix of the inputs is shown in Table 1. Note, the order of input parameters that are in the table and elements in the vector are constantly fixed for a future prediction process.

Table 1. The inputs matrix of given datasets

	Laser power	Scan speed	Powder density	Layer thickness	Pulse frequency	Hatch spacing
Tang (2003)	1	1	1	0	0	1
Morgan (2004)	0	1	0	0	1	1

Writing these inputs as input vecors:

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"A DOMAIN-DRIVEN APPROACH TO METAMODELING IN ADDITIVE MANUFACTURING."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2017, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

$$X_{tang} = [1\ 1\ 1\ 0\ 0\ 1]^T$$
$$X_{morgan} = [0\ 1\ 0\ 0\ 1\ 1]^T$$

The output vector Y is omitted in this case study since both knowledge and verification datasets have the same target output – relative density.

Three algorithms were used to characterize both input datasets. A PR model was built using the pure quadratic regression method. The Kriging model is built by ordinary Kriging method and the Gaussian correlation function with maximum likelihood approach. The ANN model is defined with 10 hidden layers. It should be noted that the candidate algorithms used in this case study are not meant to be exhaustive, but rather to represent a set of common modeling approaches. To calculate the performance, each original dataset is divided into training and testing sets with fixed ratio 80% and 20% using the Latin Hypercube based Minimum Euclidean Distance method [14]. For Tang et al (2003)'s dataset, PR works the best from three candidates as both NRSME (0.1580) and MREM (0.0220) are the lowest (Table 2). However, in the second dataset the Kriging model tested was found to be the best possible choice among the candidate algorithms. Thus, at a system level the knowledge base indicates: 1) while X=[1 1 1 0 0 1]^T, recommend model=PR; 2) while X=[0 1 0 0 1 1]^T, recommend model=Kriging. The future model selection process is built based on these rules.

Table 2. Model performance of both datasets for three candidates

	NRSME			MREM		
	PR	**Kriging**	**ANN**	**PR**	**Kriging**	**ANN**
Tang (2003)	0.1580	0.1757	0.5603	0.0220	0.0230	0.1642
Morgan (2004)	0.2018	0.1332	0.3917	0.1055	0.0669	0.1866

6.2 Modeling algorithm recommendation

The verification dataset consisted of 13 samples manufactured using the metal PBF process [49]. Compared to the datasets in the knowledge model, the experiment used carbon steel powder instead of stainless steel or a copper alloy, and has the smallest sample size (13) and number of input variables (2). The two input variables were layer thickness and hatch spacing, resulting in an input vector of:

$$X_{new} = [0\ 0\ 0\ 1\ 0\ 1]^T$$

Based on 1-NN approach, the Euclidean distance between the new and former datasets are:

$$\text{dist}(\text{new}, \text{tang}) = 2$$
$$\text{dist}(\text{new}, \text{morgan}) = 1.732$$

The knowledge model at this stage is likely insufficient due to a very limited number of example instances. Though it has these defects, the distance results show that the dataset is closer to Morgan's data than Tang's. Thus, the recommended algorithm would be a Kriging model. Once confirmed, the result can be used to update the current knowledge model with a new instance – while X=[0 0 0 1 0 1]^T, and Y is relative density, the optimal candidate model is Kriging. Once this is done, the updated knowledge model was updated and can cover the aspect of layer thickness.

For verification, the performance of each candidate model with the new dataset is shown in Table 3. All models were constructed using the same methods and model parameters as in the knowledge model datasets. Based on the performance measurement, Kriging model shows small advantages in both NRSME and MREM compared to the PR and ANN models. Thus, the result is consistent with the solution predicted using the framework.

Table 3. Model performance for the new dataset

	NRSME			MREM		
	PR	**Kriging**	**ANN**	**PR**	**Kriging**	**ANN**
Chatterjee (2003)	0.3374	0.3186	1.0448	0.0282	0.0257	0.0681

7. DISCUSSION AND FUTURE WORK

The proposed AM domain-driven metamodeling recommendation framework has the potential to provide an efficient and reliable way to predict the optimal metamodel for a new problem. It is efficient as it can avoid exhaustively testing all possible candidate algorithms once a sufficient knowledge model is constructed. Moreover, the solution can help to direct future model construction when considering data-features that might allow the user to hone in from a broad class of algorithms to a specific one. The general framework was established based on the hypothesis that certain combinations of input/output parameters have consistent behavior. The predictive solution could be made more reliable if it were derived from a larger set of consolidated knowledge. A simple demonstrative case study that included three distinct metal PBF datasets shows the algorithm prediction process.

Though the proposed framework shows a multiple of advantages in AM metamodeling problems, the details of the method need further improvement. The current set of candidate algorithms is limited; including only PR, Kriging, and ANN. While suitable for demonstration, this limited size of candidates potentially restrains higher model performance of new datasets. Furthermore, each model only has the basic modeling configuration without the ability for user modification. It may cause false results due to incomplete consideration of modeling options. For example, the finding that the ordinary Kriging model works better than a pure quadratic regression model does not mean that it also works better than a higher order PR model. Without consideration of the range of available algorithm types, the current framework may mislead the user. To overcome these disadvantages, more detailed metamodeling techniques should be added to the current framework. This work is being undertaken.

Beyond adding more broad classes of algorithms, subclasses of algorithms also need to be considered for a more robust solution. For example, consideration of different Kriging methods might enrich the study. Simple Kriging, stochastic Kriging, and dynamic Kriging may further define the Kriging class in the set of candidates. In addition to adding more

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"A DOMAIN-DRIVEN APPROACH TO METAMODELING IN ADDITIVE MANUFACTURING."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2017, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

algorithms, it may be useful to bring modeling guidelines into the framework. For example, such guidelines might indicate that ANN may not be well suited for use with small datasets. Such considerations may improve the predictive accuracy of the framework for a larger breadth of datasets.

For the specific case of AM and the PBF process, algorithms for the AM characterization process also need further improvement. There are more than 50 independent variables in metal PBF process [1, 16]. This study has included less than 1/3 of them. Another disadvantage is that the framework can only count categorical input/output vectors, rather than considering broad classes of inputs and the relative similarity between different types of variables. For example, in the review of the literature, variables relating to energy density were found to behave very similarly within a range of outputs. This is knowledge that might improve the model selection process. If included in the knowledge model, the system could possess greater insight when calculating the distance between instances. More robust parameter classification may thus be needed for more accurate prediction. Similarly, as research continues, the vectors can be further detailed and classified in multiple levels based on process knowledge and empirical data. For example, materials could be classified as 0 (single component), 0.5 (multiple components without steel), or 1 (multiple components with steel), or using some other scheme to provide greater insight into problem similarity. However, development of reasonable methods requires a more comprehensive understanding of AM processes.

All of this suggests the need for a hybrid approach that utilizes a combination of process-specific knowledge and experience, algorithmic knowledge and dataset-specific considerations. The process knowledge might consist of empirical input/output relations as in this paper, utilize knowledge of problem physics to assess the similarity of datasets and suggest several candidate classes of algorithm. Algorithmic knowledge might consist of a well defined model of broad algorithm classes and subclasses, and defined model and data attributes that affect their performance. Simple data features such as sample size and the utilized DOE methods could then be used.

The most important challenge in our research currently is a lack of data to construct a more reliable knowledge model. In the case study, the naive knowledge model consists of only two instances. To improve the model, more AM knowledge is needed. In the current knowledge model, only empirical datasets can be used to build the knowledge model. Simulation models might be used to enhance the knowledge model. For example, Ma et al (2015)'s FEA model has 10 independent AM input variables, which may allow the framework to include a broader range of problem physics [22]. Formal information models may also contribute to better knowledge construction. More candidate metamodeling algorithms should be considered into such information models to not limit the overall performance. Recent development of an AM ontology might provide the basis for more effective utilization of process specific knowledge. If this information can be utilized in a future version of the proposed framework, it can potentially boost its predictive ability and accuracy.

ACKNOWLEGEMENT

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1439683, the National Institute of Standards and Technology (NIST) under Cooperative Agreement number NIST 70NANB15H320, and industry members of the NSF Center for e-Design.

REFERENCES

[1] Gibson, I., Rosen, D.W., and Stucker, B., 2010, "Additive manufacturing technologies," Springer, .
[2] Tapia, G., and Elwany, A., 2014, "A Review on Process Monitoring and Control in Metal-Based Additive Manufacturing," Journal of Manufacturing Science and Engineering, 136(6) pp. 060801.
[3] Kamath, C., El-dasher, B., Gallegos, G. F., 2014, "Density of Additively-Manufactured, 316L SS Parts using Laser Powder-Bed Fusion at Powers Up to 400 W," The International Journal of Advanced Manufacturing Technology, 74(1-4) pp. 65-78.
[4] Wang, Z., Guan, K., Gao, M., 2012, "The Microstructure and Mechanical Properties of Deposited-IN718 by Selective Laser Melting," Journal of Alloys and Compounds, 513pp. 518-523.
[5] Huang, Y., Leu, M. C., Mazumder, J., 2015, "Additive Manufacturing: Current State, Future Potential, Gaps and Needs, and Recommendations," Journal of Manufacturing Science and Engineering, 137(1) pp. 014001.
[6] Yang, L., Peng, X., and Wang, B., 2001, "Numerical Modeling and Experimental Investigation on the Characteristics of Molten Pool during Laser Processing," International Journal of Heat and Mass Transfer, 44(23) pp. 4465-4473.
[7] Witherell, P., Feng, S., Simpson, T. W., 2014, "Toward Metamodels for Composable and Reusable Additive Manufacturing Process Models," Journal of Manufacturing Science and Engineering, 136(6) pp. 061025.
[8] Wang, G. G., and Shan, S., 2007, "Review of Metamodeling Techniques in Support of Engineering Design Optimization," Journal of Mechanical Design, 129(4) pp. 370-380.
[9] Jin, R., Du, X., and Chen, W., 2003, "The use of Metamodeling Techniques for Optimization Under Uncertainty," Structural and Multidisciplinary Optimization, 25(2) pp. 99-116.
[10] Hoffman, F. O., and Hammonds, J. S., 1994, "Propagation of Uncertainty in Risk Assessments: The Need to Distinguish between Uncertainty due to Lack of Knowledge and Uncertainty due to Variability," Risk Analysis, 14(5) pp. 707-712.
[11] Kothari, C.R., 2004, "Research methodology: Methods and techniques," New Age International, .
[12] Holman, J.P., and Gajda, W.J., 1994, "Experimental methods for engineers," McGraw-Hill New York, .
[13] Cui, C., 2016, Building Energy Modeling: A Data-Driven Approach, .
[14] Yang, Z., Eddy, D., Krishnamurty, S., 2016, "Investigating Predictive Metamodeling for Additive Manufacturing," ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Anonymous American Society of Mechanical Engineers, pp. V01AT02A020-V01AT02A020.
[15] Rice, J. R., 1976, "The Algorithm Selection Problem," Advances in Computers, 15pp. 65-118.
[16] Frazier, W. E., 2014, "Metal Additive Manufacturing: A Review," Journal of Materials Engineering and Performance, 23(6) pp. 1917-1928.
[17] Olakanmi, E., Cochrane, R., and Dalgarno, K., 2015, "A Review on Selective Laser Sintering/Melting (SLS/SLM) of Aluminium Alloy Powders: Processing, Microstructure, and Properties," Progress in Materials Science, 74pp. 401-477.
[18] Xiao, B., and Zhang, Y., 2007, "Marangoni and Buoyancy Effects on Direct Metal Laser Sintering with a Moving Laser Beam," Numerical Heat Transfer, Part A: Applications, 51(8) pp. 715-733.

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"A DOMAIN-DRIVEN APPROACH TO METAMODELING IN ADDITIVE MANUFACTURING."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2017, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

[19] Devesse, W., De Baere, D., and Guillaume, P., 2014, "The Isotherm Migration Method in Spherical Coordinates with a Moving Heat Source," International Journal of Heat and Mass Transfer, 75pp. 726-735.
[20] Roberts, I., Wang, C., Esterlein, R., 2009, "A Three-Dimensional Finite Element Analysis of the Temperature Field during Laser Melting of Metal Powders in Additive Layer Manufacturing," International Journal of Machine Tools and Manufacture, 49(12) pp. 916-923.
[21] Contuzzi, N., Campanelli, S., and Ludovico, A., 2011, "3 D Finite Element Analysis in the Selective Laser Melting Process," International Journal of Simulation Modelling, 10(3) pp. 113-121.
[22] Ma, L., Fong, J., Lane, B., 2015, "Using design of experiments in finite element modeling to identify critical variables for laser powder bed fusion," International Solid Freeform Fabrication Symposium, Anonymous Laboratory for Freeform Fabrication and the University of Texas Austin, TX, USA, .
[23] Palmer, K. D., 1998, Data Collection Plans and Meta Models for Chemical Process Flowsheet Simulators, .
[24] Siah, E. S., Sasena, M., Volakis, J. L., 2004, "Fast Parameter Optimization of Large-Scale Electromagnetic Objects using DIRECT with Kriging Metamodeling," IEEE Transactions on Microwave Theory and Techniques, 52(1) pp. 276-285.
[25] Shao, T., 2007, "Toward a structured approach to simulation-based engineering design under uncertainty," ProQuest, .
[26] Shao, T., and Krishnamurty, S., 2008, "A Clustering-Based Surrogate Model Updating Approach to Simulation-Based Engineering Design," Journal of Mechanical Design, 130(4) pp. 041101.
[27] Zhao, D., and Xue, D., 2011, "A Multi-Surrogate Approximation Method for Metamodeling," Engineering with Computers, 27(2) pp. 139-153.
[28] Bauchau, O., and Craig, J., 2009, "Structural Analysis,"Springer, pp. 173-221.
[29] Gong, H., Rafi, K., Gu, H., 2014, "Analysis of Defect Generation in Ti–6Al–4V Parts made using Powder Bed Fusion Additive Manufacturing Processes," Additive Manufacturing, 1pp. 87-98.
[30] Raghunath, N., and Pandey, P. M., 2007, "Improving Accuracy through Shrinkage Modelling by using Taguchi Method in Selective Laser Sintering," International Journal of Machine Tools and Manufacture, 47(6) pp. 985-995.
[31] Tang, Y., Loh, H., Wong, Y., 2003, "Direct Laser Sintering of a Copper-Based Alloy for Creating Three-Dimensional Metal Parts," Journal of Materials Processing Technology, 140(1) pp. 368-372.
[32] Kempen, K., Yasa, E., Thijs, L., 2011, "Microstructure and Mechanical Properties of Selective Laser Melted 18Ni-300 Steel," Physics Procedia, 12pp. 255-263.
[33] Louvis, E., Fox, P., and Sutcliffe, C. J., 2011, "Selective Laser Melting of Aluminium Components," Journal of Materials Processing Technology, 211(2) pp. 275-284.
[34] Kruth, J., Wang, X., Laoui, T., 2003, "Lasers and Materials in Selective Laser Sintering," Assembly Automation, 23(4) pp. 357-371.
[35] Morgan, R., Sutcliffe, C., and O'neill, W., 2004, "Density Analysis of Direct Metal Laser Re-Melted 316L Stainless Steel Cubic Primitives," Journal of Materials Science, 39(4) pp. 1195-1205.
[36] Beaman, J. J., Barlow, J. W., Bourell, D. L., 1997, "Solid Freeform Fabrication: A New Direction in Manufacturing," Kluwer Academic Publishers, Norwell, MA, 2061pp. 25-49.
[37] Meier, H., and Haberland, C., 2008, "Experimental Studies on Selective Laser Melting of Metallic Parts," Materialwissenschaft Und Werkstofftechnik, 39(9) pp. 665-670.
[38] Ciurana, J., Hernandez, L., and Delgado, J., 2013, "Energy Density Analysis on Single Tracks Formed by Selective Laser Melting with CoCrMo Powder Material," The International Journal of Advanced Manufacturing Technology, 68(5-8) pp. 1103-1110.
[39] Mills, K.C., 2002, "Recommended values of thermophysical properties for selected commercial alloys," Woodhead Publishing, .
[40] Box, G.E., and Draper, N.R., 1987, "Empirical model-building and response surfaces," Wiley New York, .
[41] Simpson, T. W., Booker, A. J., Ghosh, D., 2004, "Approximation Methods in Multidisciplinary Analysis and Optimization: A Panel Discussion," Structural and Multidisciplinary Optimization, 27(5) pp. 302-313.
[42] Shan, S., and Wang, G. G., 2010, "Survey of Modeling and Optimization Strategies to Solve High-Dimensional Design Problems with Computationally-Expensive Black-Box Functions," Structural and Multidisciplinary Optimization, 41(2) pp. 219-241.
[43] Cressie, N., 2015, "Statistics for spatial data," John Wiley & Sons, .
[44] Sacks, J., Welch, W. J., Mitchell, T. J., 1989, "Design and Analysis of Computer Experiments," Statistical Science, pp. 409-423.
[45] Rosenblatt, F., 1958, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." Psychological Review, 65(6) pp. 386.
[46] Yegnanarayana, B., 2009, "Artificial neural networks," PHI Learning Pvt. Ltd., .
[47] Shanmuganathan, S., and Samarasinghe, S., 2016, "Artificial neural network modelling," Springer, .
[48] Narendra, P. M., and Fukunaga, K., 1977, "A Branch and Bound Algorithm for Feature Subset Selection," IEEE Transactions on Computers, 26(9) pp. 917-922.
[49] Chatterjee, A., Kumar, S., Saha, P., 2003, "An Experimental Design Approach to Selective Laser Sintering of Low Carbon Steel," Journal of Materials Processing Technology, 136(1) pp. 151-157.

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"A DOMAIN-DRIVEN APPROACH TO METAMODELING IN ADDITIVE MANUFACTURING."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2017, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

Proceedings of the ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2017
August 6-9, 2017, Cleveland, Ohio, USA

DETC2017-67794

INVESTIGATING GREY-BOX MODELING FOR PREDICTIVE ANALYTICS IN SMART MANUFACTURING

Zhuo Yang, Douglas Eddy, Sundar Krishnamurty, Ian Grosse
University of Massachusetts Amherst
Department of Mechanical and Industrial Engineering
Amherst, MA 01003
Email: [zhuoyang, dceddy]@engin.umass.edu, [skrishna, grosse]@ecs.umass.edu

Peter Denno, Yan Lu, Paul Witherell
National Institute of Standards and Technology
Engineering Laboratory
Gaithersburg, MD 20899
Email: [peter.denno, yan.lu, paul.witherell]@nist.gov

ABSTRACT

This paper develops a two-stage grey-box modeling approach that combines manufacturing knowledge-based (white-box) models with statistical (black-box) metamodels to improve model reusability and predictability. A white-box model can use various types of existing knowledge such as physical theory, high fidelity simulation or empirical data to build the foundation of the general model. The residual between a white-box prediction and empirical data can be represented with a black-box model. The combination of the white-box and black-box models provides the parallel hybrid structure of a grey-box. For any new point prediction, the estimated residual from the black-box is combined with white-box knowledge to produce the final grey-box solution. This approach was developed for use with manufacturing processes, and applied to a powder bed fusion additive manufacturing process. It can be applied in other common modeling scenarios. Two illustrative case studies are brought into the work to test this grey-box modeling approach; first for pure mathematical rigor and second for manufacturing specifically. The results of the case studies suggest that the use of grey-box models can lower predictive errors. Moreover, the resulting black-box model that represents any residual is a usable, accurate metamodel.

1. INTRODUCTION

Smart manufacturing is becoming increasing possible as access to technology improves. Industry is, and will continue to be, increasingly reliant on data and predictive analytics to improve overall process efficiencies [1]. With this trend, industry is now collecting data at never before seen rates in hopes of gaining competitive advantages related to their products and processes. Often data are collected without regard for their interrelation, and it is not readily apparent how the collected information can be used to improve system efficiencies. To address this issue, we investigate a novel metamodeling technique based on the context from which the data was acquired and the domain in which it is relevant.

White-box modeling methods use knowledge such as rules and theories, to formulate models such as those that represent physical phenomena. Such classical white box modeling methods have been used for thousands of years. Newton's Law or Euler–Bernoulli bending theory [2] is a classic example of a traditional physical model. Such models usually require comprehensive knowledge of the target system and are usually represented by parametric formulation. For instance, in manufacturing, physics-based models are often derived from theoretical analysis that mostly focuses on individual sub-processes with idealized assumptions. In reality, the actual multi-physical system of manufacturing may involve numerous interactions among these sub-processes. Such complexity can be difficult to fully understand. For example, in additive manufacturing (AM), the isotherm migration method develops a thermal model to calculate the temperature on a powder surface being heated by a laser beam modeled as a point source

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2018, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

[3]. However, the complex inter-relationship between parameters of Powder Bed Fusion (PBF) processes renders these theoretical analyses insufficient for the needs of many practical applications [4].

Metamodels, also known as surrogate models, are statistical models that use a "black-box" approach to represent unknown systems and hence do not require detailed knowledge of the underlying physical phenomena [5]. Metamodeling focuses on the input/output parametric values while ignoring the complex inter-relationships within the unknown system (i.e. the black-box, which statistically approximates the relationships based purely upon data values). Rather than incorporating any physical knowledge, the predictability of metamodels completely relies upon statistical features such as sampling strategy and modeling algorithm. Metamodels can optimally reduce the inaccuracies that arise from incomplete knowledge. For example, an adaptive sampling method that iteratively updates the metamodel with a new sample data point can gradually refine the model's predictive power [6, 7].

Furthermore, predictability or efficiency of a metamodel can be improved by selecting the most suitable algorithm [8] and/or best combination of algorithms [9] in cases where it is not possible to acquire more data points.in the desired locations in the design space. Advanced modeling algorithms such as a dynamic Kriging method and artificial neural networks (ANN) [10] can significantly improve model efficiency and predictability for these types of inflexible datasets. However, such metamodels built by pure statistical approaches usually lack information about the model's physical meaning and assumptions due to a large degree of data-dependency. Moreover, modeling inaccuracies might accumulate during the model construction process due to the lack of physical knowledge about the critical features of the represented system [5]. Thus, both white and black box approaches alone have accuracy limitations due to different reasons.

Due to these intrinsic limitations in both approaches, a technique that can harness the advantages of both white and black box models while reducing their disadvantages is desirable for complex problems with understood subdomains. The modeling approach known as grey-box, or hybrid modeling, was invented to combine the benefits of domain knowledge and empirical information [11]. The models generated by this approach can obey general physical rules (white box) while optimizing the parameters from actual experimental data (black box).. Many of the newer and less established white box manufacturing physics-based models and numerical simulations may be founded on incomplete and/or inaccurate knowledge and idealized assumptions. For example, the previously mentioned isotherm migration model in AM does not account for the influence of powder particle size, part geometry, and environmental conditions [3]. The calculated solutions from current AM physics-based models are usually limited in the scope of what they describe, diminishing their predictive capability. Though AM metamodels can potentially avoid these errors, they usually require a large number of expensive samples and may not be reusable. Many examples in smart manufacturing have similar modeling challenges. These barriers potentially limit the adaptability of metamodeling in any manufacturing domain. Thus, neither approach can optimally construct robust, usable manufacturing models alone.

This paper aims to develop a grey-box modeling approach which combines the benefits of traditional serial methods (where black and white box knowledge is applied sequentially) and parallel methods (where knowledge from black and white boxes is composed before being applied). The result is a hybrid combination of knowledge of physical phenomena and statistical information. To address the challenge of combining knowledge of physical phenomena with statistical information, a two-stage approach is used. The first stage deploys a serial grey–box approach to build a statistical black-box model to estimate the errors caused by the inaccuracies in a white-box. The second stage uses the model from the first stage to estimate the basic solution and the residual solution. The final solution is a combination of these two.

Section 2 and 3 provide fundamental background knowledge relating to metamodeling and grey-box modeling techniques. Section 4 introduces this general algorithm of a two-stage grey-box modeling approach for additive or smart manufacturing. Case studies using a general mathematical example and a representative metal PBF AM problem are presented in Section 5. Section 6 discusses the results and identifies future work for this study.

2. OVERVIEW OF METAMODELING TECHNIQUES

This section briefly reviews the metamodeling techniques used in this work. Polynomial regression (PR) and Kriging methods were investigated in this study as they cover the spectrum of both parametric and non-parametric modeling algorithms [12]. PR is popular in that its model is represented by parametric formulation. The Kriging method, on the other hand, is an interpolation approach that uses positioning information for data estimation instead of conventional mathematical formulation. The general mathematical formulation of any metamodel can be expressed as:

$$y(\tilde{x}) = f(\tilde{x}) + \varepsilon \quad (1)$$

where $y(\tilde{x})$ represents the actual output for new point $\tilde{x}$ [13]. $f(\tilde{x})$ is a known function derived statistically from data that produces the model estimate as a function of $\tilde{x}$, ε is any residual error, and $\tilde{x}$ represents the set of the independent input variables. For different modeling methods, the composition of each of these elements could be different.

2.1 Polynomial Regression

PR is a higher order variation of linear regression in which an n^{th} order polynomial is used to formulate the relationship between the independent variables x and the outcome y [13]. It is popular in various engineering domains due to its efficiency. A second order quadratic polynomial function would have the form:

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2018, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

$$\hat{y} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i=1}^{k} \sum_{j=1}^{k} \beta_{ij} x_i x_j \quad (2)$$

where β_0, β_i, and β_{ij} are regression coefficients, and k is the number of design variables.

2.2 Kriging

Unlike parametric techniques that produce an actual formulation. Kriging methods are non-parametric methods that build their estimation based on the position of the samples. The basic assumption in kriging is that the estimating point (unknown point) can be represented by observed points (known points) based on spatial correlation [14]. The estimation process is completed by a variogram or so called spatial correlation functions [15, 16]. The general form of kriging estimation for an unknown predicted value of a point Z_E for a single outcome is:

$$Z_E = \bar{Z} + \sum_{i=1}^{n} \lambda_i (Z_i - \bar{Z}) \quad (3)$$

where $\bar{Z}$ represents the regional mean value of the response and λ_i is the distance-correlated weight value, which is determined by the computation of spatial correlation.

To approach the weight value, one should first compute the spatial correlation R between data points. The value of spatial correlation can be derived from:

$$R(\theta, x_i, x_j) = \prod_{l=1}^{n} \exp(-\theta_l (x_{i,l} - x_{j,l})^2) \quad (4)$$

where $x_{i,l}$ is the lth component of the i^{th} vector x_i [17]. $R(\theta, x_i, x_j)$ depends upon the location of points x_i and x_j, and the correlation parameter, θ. Kriging methods have multiple forms such as simple kriging, ordinary kriging, regression kriging, etc.[16]. In this paper, results from kriging models are obtained from the ordinary kriging method.

The correlation matrix can then be formulated. A problem with n given data points can be presented as:

$$\mathbf{C} = \begin{bmatrix} R(\theta, x_1, x_1) & R(\theta, x_1, x_2) & \dots & R(\theta, x_1, x_n) \\ R(\theta, x_2, x_1) & R(\theta, x_2, x_2) & \dots & R(\theta, x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ R(\theta, x_n, x_1) & R(\theta, x_n, x_2) & \dots & R(\theta, x_n, x_n) \end{bmatrix} \quad (5)$$

Similarly, the correlation vector B that presents the correlation between the new point x_E and all given points is formulated as:

$$\mathbf{B} = \begin{bmatrix} R(\theta, x_1, x_E) \\ R(\theta, x_2, x_E) \\ \vdots \\ R(\theta, x_n, x_E) \end{bmatrix} \quad (6)$$

The weight vector $\Lambda = [\lambda_1, \lambda_2, \dots \lambda_n]^T$ can now be calculated by C and B:

$$\Lambda = C^{-1} B \quad (7)$$

3. OVERVIEW OF GREY-BOX MODELS

A grey-box model is a hybrid model that combines different types of models such as physics-based models, numerical simulation models and statistical models [18]. The term "grey-box" stems from the mixture of white-box and black-box models. A conventional grey-box model uses a physical formulation to maintain the physical interpretation and uses data to estimate parameters [18]. In general, the basic structure of a grey-box model is inherited from knowledge and further improved by experimental data.

Grey-box model development can be summarized into three steps: 1) construct the foundation for the system with a simplified knowledge model; 2) determine the physical parameters from the description of the system behavior; 3) identify the value of model parameters from actual data [8]. The relationship between these three types of model and knowledge sources is shown in Figure 1. The proposed grey-box metamodeling method was developed based upon this viewpoint.

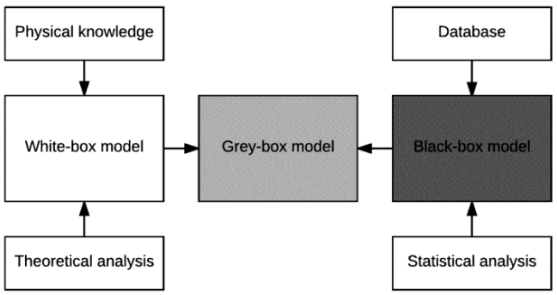


Figure 1. Relationships among physics-based white-box, statistics-based black-box, hybrid grey-box models and knowledge sources

Grey-box models can be generally classified into serial approach and parallel approaches [19, 20], which are shown in Figure 2. A serial approach aims to sequentially fill the gap between knowledge and experimental data. For example, the uncertainties raised from incomplete knowledge of a white-box model can be reduced by accompanying that model with actual data. A parallel approach, alternatively, aims to use both models

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2018, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

together to estimate the correct results that would be difficult to approach by either a white or black-box model individually. A grey-box model with serial structure focuses on reducing the error between the prediction from physical model and actual result from experiments. For example, Duarte et al. [21] developed a hybrid modeling approach that combined knowledge and mechanistic, rather than statistical models, to improve traditional model performance. In this approach, the first model is built based on first-principles system behavior, and the second model estimates the residuals between real data and mechanistic predictions.

A traditional grey-box model with parallel structure uses data to estimate the correct values of model responses which are difficult to approach given incomplete knowledge of phenomena [22]. For example, Psichogios et al. [19] hybrid neural network model utilizes a partial first principles model. This modeling approach combines available prior knowledge with an artificial neural network (ANN) to derive an estimator of unmeasured process parameters. This hybrid structure can interpolate and extrapolate much more accurately than a standard "black-box" ANN with significantly fewer training sample points to accompany the knowledge model.

The next section introduces the two-stage grey-box modeling approach developed for manufacturing problems in this study. To get the final prediction, the data serially flows into both types of grey-box models for the purpose of constructing the black-box model and estimating the residual. To demonstrate this approach, we chose a complex manufacturing process that we believe could particularly benefit. The AM-specific grey-box modeling approach that is introduced next in Section 4 is built upon both the Type II serial approach and parallel approach shown in Figure 2 based on current AM challenges.

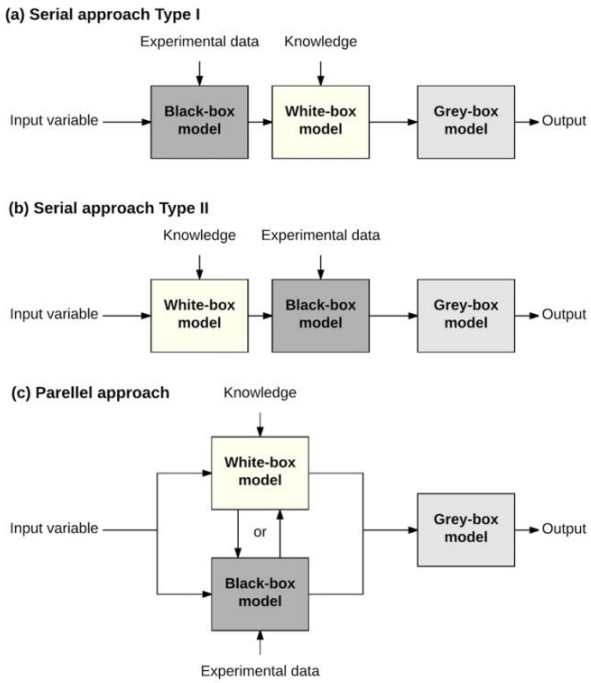


Figure 2. Basic grey-box modeling approaches

4. GREY-BOX METAMODELING FOR AM

Additive manufacturing (AM) and other smart manufacturing techniques are being widely used in various domains such as aerospace [23], medical devices [24] and energy systems [25]. However, many barriers and challenges, such as the large uncertainty of AM process results, have prevented its further adoption in industry [25]. The relationship between process parameters and mechanical properties are not fully understood for AM processes. For example, relative density, one of the major structural properties of the parts produced by metal PBF processes, depends upon multiple AM parameters such as laser power, scan speed, pulse frequency, and layer thickness [26]. Previous studies show that a typical metal PBF process consists of four general sub-systems classified by related physical phenomenon. Each sub-system can be further divided into multiple sub-processes [27]. A general AM process can involve more than fifty independent parameters [4]. For example, the melt pool sub-system is related to a number of factors that involve both thermal and fluid mechanics [28]. Though difficult, AM models built upon theoretical analysis, numerical simulation and statistical modeling have been developed for predictive purposes in recent years [7].

The general procedure to construct an AM grey-box model by using the approach introduced in this paper is shown in Figure 3 and Figure 4. First, the method builds the white-box model from available prior knowledge. If knowledge was derived from theoretical analysis, the white-box model can be directly represented by a parametric formulation. Alternatively,

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2018, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

a parametric model can be derived through an approximation of a physics-based model using a formulation such as FEA for computational simplification. However, if the knowledge is based on a complex numerical simulation that requires high computational cost, the white-box model can be redesigned to be represented by some simplified parametric function such as a PR response surface model to more rapidly estimate a white box model using fewer sample points of that expensive data. The sampling data to construct that PR model could be collected by a technique such as space filling sampling (SFS) or sequential infilling sampling (SIS) [6] from data generated by an adequate number of simulations. It can be expected that the solution from the white-box model would contain large errors due to limited knowledge.

The next step is to build a black-box model from additional information. Potential sources of the additional information could be actual experimental data or a higher fidelity simulation used to generate the data. The black box model captures both the discrepancy between a lower fidelity FEA type of model and the real process as well as the discrepancy between a high fidelity model (FEA model) and the simplified model. The input values of this additional data are entered into the constructed white-box model to calculate the corresponding output responses. This computational output from the white-box model is next compared with the actual output values of the additional data to calculate each difference. This difference can be considered the white box model's actual residual at that data location as shown in Figure 3. Since each pair of the computational output and the actual output has the same input variables, the residual value directly represents the accumulated errors caused by incomplete and/or imperfect knowledge used to construct the white box model. The black-box model is used here to evaluate the relation between input variables and the estimated residual.

At this first stage, the serial grey-box structure is established based on the type II serial approach that is shown in Figure 2b: the output from the white-box becomes an intermediate input to the black-box. This serial grey-box approach is used to build a black-box model to estimate the residual value that cannot be derived from the white-box alone. That residual is the difference between the responses predicted by the white and black boxes. The inputs to the black box are those used to generate the white-box responses. The black-box model uses the kriging method to model the relationship between input variable values and residual response values. The kriging method is applied since this interpolation approach helps to avoid any significant intrinsic error in the resulting model [29]. Once the black-box is created, it can compute the estimated residual for any new data point. The approach, illustrated in Figure 3, establishes the black-box model used to derive the grey-box model created in the subsequent steps shown in Figure 4.

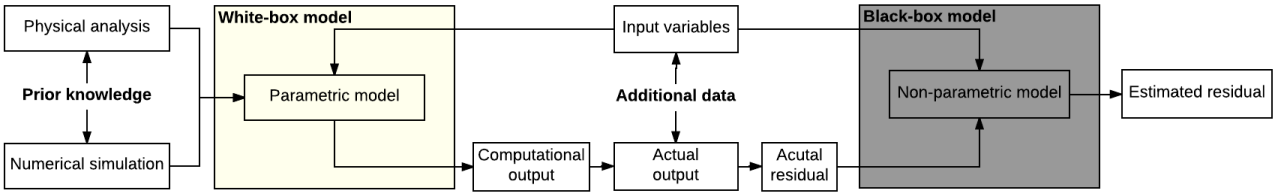


Figure 3. General workflow of the first stage

Figure 4 depicts the second stage of the process, wherein the white-box and black-box of the residual built in the first stage are composed to a parallel structure. It is considered a parallel structure because the given values of input variables are entered into white-box and black-box models simultaneously. The white-box in Figure 4 is the same as the one in Figure 3. However, at this grey-box modeling stage, the output from the white-box directly estimates the final solution in concert with the estimated residual from the black-box. For each new data point prediction, the output from the white-box is used as the basic solution. The residual solution from the black-box is the estimated residual for that same new data point. The final solution is the combination of basic solution with its estimated residual, or the results from both stages.

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2018, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

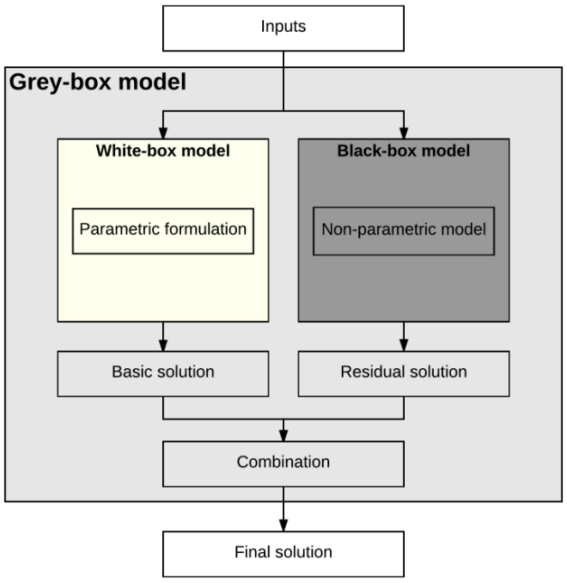


Figure 4. General workflow of the second stage

To illustrate the proposed grey-box modeling method, two case studies are presented in the next section. The first, a classical mathematical example of a mystery function [5, 30] demonstrates the process of constructing grey-box models from pre-existing knowledge that can be expressed numerically. The second example illustrates the use of this grey-box modeling technique to predict the relative density resulting from an AM process and represented by actual experimental data.

5. ILLUSTRATIVE CASE STUDIES

To illustrate the method introduced in the prior section, two case studies are presented in this section. The mystery function examples in Section 5.1 illustrate the process of grey-box model construction for different types of knowledge.

Maximum relative error magnitude (MREM) and average relative error magnitude (AREM) are used to represent the model predictability [7]. These two metrics are used to evaluate metamodel predictability because the combination of both metrics reveals both the overall predictability and the worst case of predictability of a metamodel in its design space. The formulation of MREM and AREM are:

$$MREM = \max\left(\frac{|y_i - \hat{y}_i|}{y_i}\right) \quad (y_i \neq 0) \tag{8}$$

$$AREM = \frac{1}{m}\left(\frac{\sum_{i=1}^{m}|y_i - \hat{y}_i|}{y_i}\right) \quad (y_i \neq 0) \tag{9}$$

where y_i is the observed value from given data, $\hat{y}$ is the value predicted by the metamodel of the data points that were not selected to construct the metamodel, and m is the number of data points.

5.1 Case studies: mystery function problem

A classical mystery function [5, 30] is brought into this study to mimic a complex unknown system. The function $f(x_1, x_2)$ that represents a nonlinear and complex system is used to generate experimental results used for model creation and assessment. The original equation of this mystery function is:

$$Y = f(x_1, x_2) = 2 + 0.01(x_2 - x_1^2)^2 + (1 - x_1) + 2(2 - x_2)^2 + 7\sin\left(\frac{x_1}{2}\right)\sin\left(\frac{7x_1x_2}{10}\right) \tag{10}$$

x_1 and x_2 are two input variables and Y is the actual output. The true surface and contour plots of the original mystery function are shown in Figure 5. To illustrate the effectiveness of this method in Section 5.1.1, the original equation is manipulated to illustrate a scenario similar to that of an inaccurate white-box model representing model construction with incomplete prior knowledge. In this situation, a parametric formulation is accessible before constructing the grey-box. However, Section 5.1.2 simulates a situation where the parametric white-box model cannot be directly derived from current knowledge. In that case, the prior knowledge was delivered by running a hypothetical simulation-based model, i.e. the manipulated function $f_k(x_1, x_2)$. These two examples illustrate how to use this grey-box modeling approach for different types of problems.

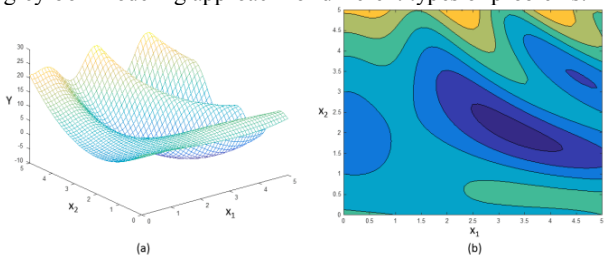


Figure 5. (a) True 3D surface plot and (b) contour plot of the original mystery function

5.1.1 Case study: theoretical physics-based formulation

In this example, the available prior knowledge is assumed derived from theoretical analysis and represented by an inaccurate parametric formulation. In this paper, it is assumed that the white-box models are reasonable representations of the manufacturing phenomena being modeled, and no validation step was included in our approach. Thus, to mimic this condition, the original mystery function is manually manipulated to represent a white-box model as:

$$\tilde{Y} = f_k(x_1, x_2) = 2 + 0.01(x_1 - x_1^2)^2 + (1 - x_1) + 2(2 - x_2)^2 + 7\sin\left(\frac{x_1}{2}\right)\sin\left(\frac{5x_1x_2}{10}\right) - 0.4x_1\sin(2x_1)\cos(x_2) \tag{11}$$

where subscript k indicates a function derived from knowledge. $\tilde{Y}$ is computational output from the white-box model $f_k(x_1, x_2)$. Plots of the white-box model are shown Figure 6 (earlier stage of grey-box modeling). After the manipulation, the 3D surface

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2018, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

maintains its general shape but several characteristics are changed, which can be observed in the figure. For example, the original local minima and maxima have shifted and the original sharp ridges became flatter. These changes result from the inaccurate white-box model. If we use the correct data from the original function to test current white-box model, the MREM and AREM are equal to 942.43 and 2.74, respectively. The large error indicates that the white-box model has very low fidelity and large predictive errors.

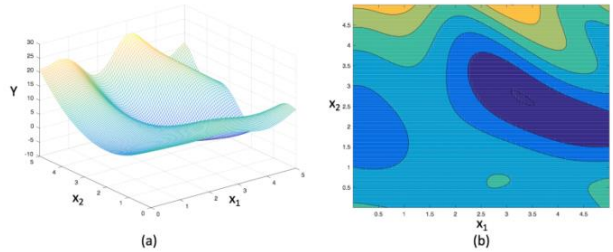


Figure 6. (a) 3D surface and (b) contour plots of manipulated mystery function

Though the white-box model has defects, it can contribute to a grey-box model. As mentioned in the last paragraph, the general information delivered from this white-box model is a reasonable representation of the level of knowledge to be expected from a white-box since the plots are generally similar to its original shape in that the local optima are still located close to their original positions. The next step is to add additional information to the initial, low fidelity prediction obtained from the white-box model. This high fidelity data is used alongside the low fidelity white-box prediction to build the black-box model. The additional information was generated from the original function using Latin Hypercube Sampling (LHS) [31] to generate 100 new data points. These additional data points represent an experimental or high fidelity model result as they were generated from the original function f(x_1, x_2), which is defined as a high fidelity system without significant error.

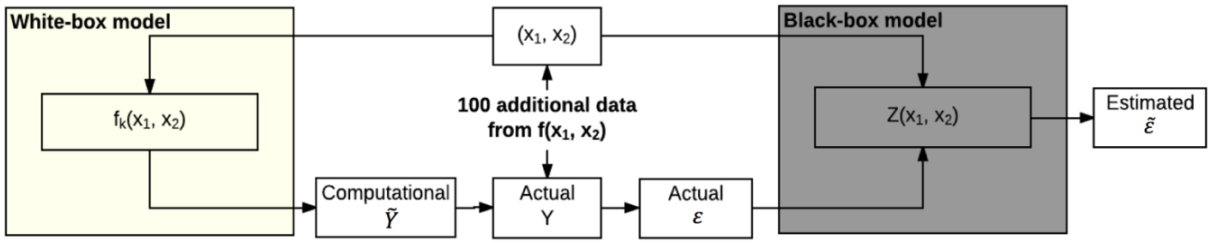


Figure 7. Black-box model construction to estimate residual

Figure 7 shows the process to construct the black-box model using the Type II serial approach (second stage of grey-box modeling). The input variables x_1 and x_2 are first entered into the white-box model $f_k(x_1, x_2)$ and used to calculate the computational output $\tilde{Y}$. The residual ε is the difference between $\tilde{Y}$ and actual output Y. The input variables (x_1, x_2) and the residual ε are next used to construct the black-box model $Z(x_1, x_2)$ using the Ordinary Kriging method. The black box model is used to compute the estimated error $\tilde{\varepsilon}$ in subsequent steps. Table 1 shows some examples from the 100 data points for illustration of the process shown in Figure 8. For example, one of the additional data points (3.73, 0.98) has an actual output 6.1881. This input when entered into the white-box model yields a prediction of 8.1416. The actual residual ε is next derived based on $\varepsilon = Y - \tilde{Y}$, which is equal to -1.9535. Once the black-box model is built from the residual values, it can estimate the residual of any unknown point from its input variable values. This estimated residual represents an expected difference between the white-box prediction and an unknown actual output. As a result, the final grey-box solution $\tilde{Y}_{final}$ should at any point be equal to $\tilde{Y} - \tilde{\varepsilon}$. This value combines the results from both stages, as shown in Figure 8. The white-box model in the parallel grey-box structure is the same one used in the prior serial approach. For any new point, the grey-box would combine basic solution $\tilde{Y}_{new}$ and the estimated residual $\tilde{\varepsilon}_{new}$ to get the final solution at that point location.

Table 1. Results at some sample data point locations

Input (x_1, x_2)	Actual output (y)	Computational output ($\tilde{Y}$)	Actual residual (ε)
(3.73, 0.98)	6.1881	8.1416	-1.9535
(4.48, 4.03)	9.3751	11.9010	-2.5258
(0.23, 1.83)	3.0593	3.0066	0.0527
(4.33, 2.98)	5.0025	4.7074	0.2951

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2018, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

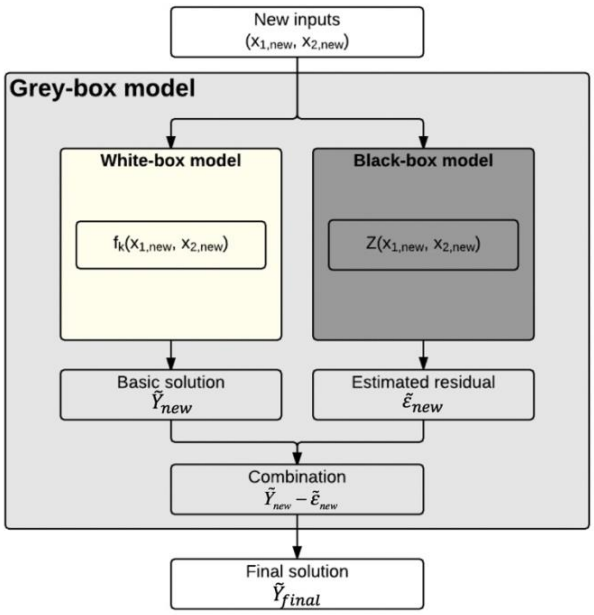


Figure 8. Grey-box model construction

One thousand randomly generated data points from the original mystery function were used to validate the resulting grey-box model. The grey-box model has reduced the initial white-box MREM from 942.24 to 2.7452 and AREM from 2.74 to 0.0359. The 3D surface and contour plots shown in Figure 9 are significantly improved and very close to the true plots of the original mystery function (Figure 5).

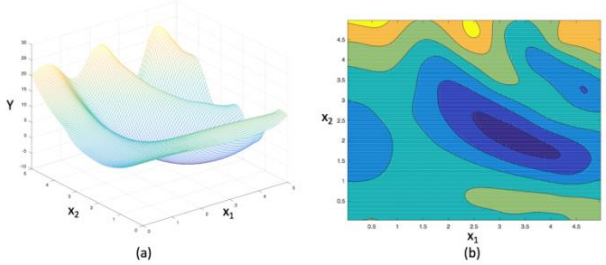


Figure 9. (a) 3D surface and (b) contour plots for grey-box model built based on 100 additional data points.

The grey-box constructed with 100 additional data points improved the initial white-box model globally. However, the MREM which represents the local error of the model remains higher than expected. It may be that the information provided by the additional dataset is insufficient. To further test the proposed method, Table 2 shows the MREM and AREM of grey-box models that are constructed with different numbers of additional data points. The top number of zero data points is a case where no data is available and the results are derived from the white-box model only. As shown, the model performance can decrease exponentially with more points in this process. Convergence criteria can be established to determine the desired accuracy.

Table 2. Model performance of additional quantity of data

Number of additional data	MREM	AREM
0	942.4327	2.7451
100	2.7452	0.0359
200	0.9372	0.0030
500	0.0019	0.0001

5.1.2. Case study: simulation-based knowledge

Many times, a theoretical physics-based parametric model is hard to access for complex problems. Simulation-based models have become more and more popular as basic reference points. Here, the initial knowledge-based parametric model f_k is assumed to be no longer available. Instead, a hypothetical simulation model replaces the former parametric white-box model. As a result, the function $f_k(x_1, x_2)$ cannot be used to generate the data needed to directly construct a black-box model and a subsequent grey-box model. Thus, a simplified white-box model is necessary since it is costly to run a high fidelity simulation for each point. To address this issue, a PR model was built to represent the white-box model. The manipulated function in Section 5.1.1 was assumed to be the simulation model. 1000 simulated data points were generated from function $f_k(x_1, x_2)$ and were used to create the PR model using LHS. The reason that the manipulated function f_k was employed instead of directly using the original mystery function is because the simulation-based model is also assumed to be low fidelity. The white-box model in PR form was generated as:

$$\tilde{Y}_{PR} = f_{PR}(x_1, x_2) = -0.3048 - 2.753x_1 - 0.228x_2 + 3.543x_1^2 + 1.973x_1x_2 + 8.582x_2 + 0.5552x_1^3 + 0.9604x_1^2x_2 + 0.9191x_1x_2^2 + 0.6433x_2^3 - 1.103x_1^4 - 0.7201x_1^3x_2 - 0.3864x_1^2x_2^2 - 0.5393x_1x_2^3 - 1.435x_2^4 \quad (12)$$

The R^2 value of this PR model is 0.7296. Comparing the PR white-box model to the original function yields an MREM of 846.6876 and an AREM of 1.9458. This indicates that this white-box model has poor predictability. This finding is reflected visually in the 3D surface and contour plots shown in Figure 10. In this figure, the shape is completely different from the original model (Figure 5). The ridges on the original surface disappeared. Thus, it is necessary to use the additional data points to build the grey-box model.

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2018, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

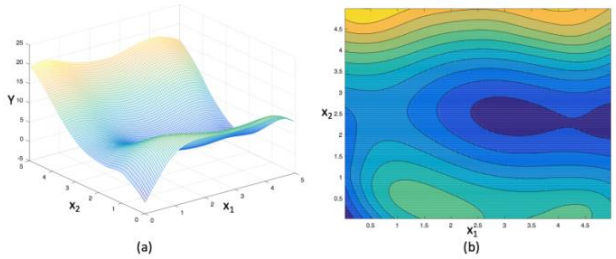


Figure 10. (a) 3D surface and (b) contour plots of the initial white-box model

The general updating process is similar to that shown in Section 5.1.1. The same 100 additional data points were used in this example compare the difference between results from Section 5.1.1 and Section 5.1.2. The Kriging black-box model was formulated by input variables and the corresponding residual values using the same procedure that was described in Section 2.2. The grey-box model was then developed by combining the PR and Kriging models by the process shown in Figure 8. The same validation process was executed to evaluate the model performance with the same validation dataset that was used in Section 5.1.1. The final MREM and AREM of this grey-box model are 3.4318 and 0.0506, which is slightly higher than using physics-based white-box model with same amount of additional data. The plots for this grey-box model are shown in Figure 2. The results from using additional data points are listed in Table 3.

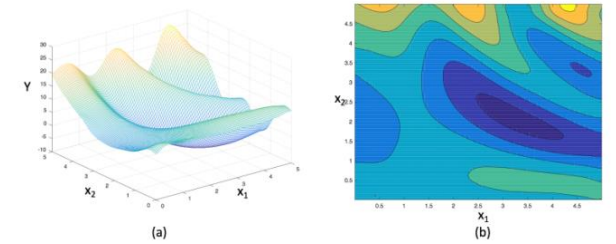


Figure 11. (a) 3D surface and (b) contour plots of the grey-box model built by simulation-based knowledge and additional actual data.

Table 3. Model performance of additional data

Number of additional data	MREM	AREM
0	846.6876	1.9458
100	3.4318	0.0506
200	1.5988	0.0055
500	0.0052	0.0001

5.2 Case study: metal PBF problem

This example uses the proposed method to build a grey-box model for a realistic PBF problem. Louvis and associates' experiments with a PBF process measured the relative density produced by different scan speed (v) and hatch spacing (d) for different aluminum alloy powders [32]. Relative density is the ratio of the actual part density to that of a completely filled solid with no porosity. The experimental results indicate higher relative density is generally produced by lower scan speed and closer hatch spacing. However, the relative density has unique behavior for specific powders and machines. For example, the density of AlSi12 and 6061 aluminum powder produced by the same parameters in different machines have different results [32], which indicates the a model built based on 6061's data may not be accurate for AlSi12. Instead of building an expensive new model, this study uses the findings from 6061's data to construct a grey-box model for AlSi12 for illustrative purposes.

In this case, there is no available physics-based knowledge to build the white-box model since the only available prior knowledge is from historical experiments. Therefore, the prior experimental knowledge of 6061 powder [32] was used to build a PR model to serve as the white-box model for illustrative purposes just as was done in Section 5.1.2. The reported measurements from AlSi12 powder were used as the additional information. First, the 177 data points from the 6061 powder experiment were used to build the PR based white-box model. The resulting quadratic model was generated as:

$$\tilde{Y}_{PR} = f_{PR}(v,d) = 81.54 + 116.83v - 0.0127d - 0.079vd - 332.39v^2 + 7082600d^2 \quad (13)$$

The initial R^2 value of this model is 0.953. The set of 36 data points from the AlSi12 PBF experiment was divided into two sets. 80% (29 points) of the data was extracted from the initial dataset to use as additional information for grey-box construction. The remaining 20% (7 points) of the data set was used to validate the models. Table 4 lists the MREM and AREM for different types of models based on the data. The 7 validation data points from the AlSi12 experiment were entered into all three types of models to evaluate and compare the predictive accuracy. The pure white-box is the PR model built using the 6061 powder experiment. The pure black-box model represents the model built with the 29 AlSi12 data points with kriging method. The grey-box represents the model built with input from both experiments using the same method presented in the prior sections. As shown, even though the pure white-box model has low predictive error, the model can be further improved by the grey-box modeling approach. The MREM of the original model is reduced from 0.0375 to 0.0238, which is a 37% improvement. Compared to the pure black-box model, the MREM of the grey-box model reduced from 0.0485 to 0.0238, which is a 51% improvement after the completion of both modeling stages.

Table 4. The performance of different types of models

	MREM	AREM
Pure white-box	0.0375	0.0170
Pure black-box	0.0485	0.0169
Grey-box	0.0238	0.0134

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2018, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

6. DISCUSSION

Effectively deploying predictive analytics in smart manufacturing is a challenge that many now face. This challenge is highlighted in AM, where current AM models often lack comprehensive information, and where information could be either knowledge-based or statistically generated. The lack of the former is typically the result of an incomplete understanding of the physical processes of AM. The lack of empirical data, on the other hand, may be caused by the difficulty of instrumenting AM processes, and more generally, the expense of producing AM parts. Even as more and more empirical data is available in the coming years, it is still difficult to duplicate all the conditions and the model predictability for all data sets. The experimental results can exhibit noticeable differences even where the experiments are operated in similar AM processes with comparable process parameters. It is also very difficult to build the connection between simulations and actual experimental data. These difficulties result from the uncertainties in and complexities of AM processes. The uncertainties significantly reduce the utility of AM predictive models since a well-validated model from one dataset may be difficult to apply to other experimental conditions.

The highlight of the two-stage grey-box modeling approach developed in this paper is that it can combine disparate knowledge and information together to produce an accurate hybrid model. To further extract the information from limited knowledge, this two-stage grey-box structure can functionally improve the predictive accuracy. In the example in Section 5.1, the original white-box model produced a large global predicative error, with an AREM over 200%. However, when updated with 100 additional data points, the grey-box approach reduced that same global error to less than 5%. These results suggest that the grey-box approach can improve the model even if the original knowledge-based model is very weak. Moreover, the results were derived from a highly nonlinear mathematical example which is even more complex than common AM problems. The 100 additional data points can sufficiently update the initial white-box model to a higher accuracy. However, a smaller sample size is expected in actual AM experiments. It is thus desirable to further reduce that sample size needed to achieve the higher model predictability.

The involved metamodeling algorithms were used as primary candidates to build the grey-box in this paper. However, the general modeling process should have no bias to other black-box modeling techniques. Any suitable algorithm that can improve the model predictability might be introduced in future work.

In grey-box modeling, the reusability of a model built on prior knowledge can be improved when combined with additional information about problem specific conditions. In Section 5.3, the case study investigated the performance of a grey-box model with two different PBF datasets generated with similar experimental conditions. The two experiments were completed in different PBF machines, with different metal powders, and unknown experimental conditions such as chamber temperature and layer thickness. Both the pure white-box and black-box models resulted in a higher MREM than the combined grey-box model when predicting the other dataset. The grey-box model reduces the MREM by about 37% compared to the white-box model and 51% compared to the black-box model. Thus, the grey-box approach provides a reliable way to reuse the knowledge from one manufacturing case to another.

ACKNOWLEGEMENT

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1439683, the National Institute of Standards and Technology (NIST) under Cooperative Agreement number NIST 70NANB15H320, and industry members of the NSF Center for e-Design.

REFERENCES

[1] Lu, Y., Witherell, P., Lopez, F., 2016, "Digital Solutions for Integrated and Collaborative Additive Manufacturing," ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V01BT02A033-V01BT02A033.

[2] Bauchau, O., and Craig, J., 2009, "Structural Analysis,"Springer, pp. 173-221.

[3] Devesse, W., De Baere, D., and Guillaume, P., 2014, "The Isotherm Migration Method in Spherical Coordinates with a Moving Heat Source," International Journal of Heat and Mass Transfer, 75pp. 726-735.

[4] Frazier, W. E., 2014, "Metal Additive Manufacturing: A Review," Journal of Materials Engineering and Performance, 23(6) pp. 1917-1928.

[5] Shao, T., 2007. Toward a structured approach to simulation-based engineering design under uncertainty. University of Massachusetts Amherst.

[6] Shao, T., and Krishnamurty, S., 2008, "A Clustering-Based Surrogate Model Updating Approach to Simulation-Based Engineering Design," Journal of Mechanical Design, 130(4) pp. 041101.

[7] Yang, Z., Eddy, D., Krishnamurty, S., 2016, "Investigating Predictive Metamodeling for Additive Manufacturing," ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V01AT02A020-V01AT02A020.

[8] Cui, C., 2016. Building Energy Modeling: A Data-Driven Approach (Doctoral dissertation, ARIZONA STATE UNIVERSITY).

[9] Zhao, D., and Xue, D., 2011, "A Multi-Surrogate Approximation Method for Metamodeling," Engineering with Computers, 27(2) pp. 139-153.

[10] Shanmuganathan, S., and Samarasinghe, S., 2016, "Artificial neural network modelling," Springer, .

[11] Kristensen, N. R., Madsen, H., and Jørgensen, S. B., 2004, "A Method for Systematic Improvement of Stochastic Grey-Box Models," Computers & Chemical Engineering, 28(8) pp. 1431-1449.

[12] Jin, R., Chen, W., and Sudjianto, A., 2004, "Analytical Metamodel-Based Global Sensitivity Analysis and Uncertainty Propagation for Robust Design," Sae Sp, 14(429) pp. 47-54.

[13] Simpson, T., Mistree, F., Korte, J. and Mauery, T., 1998, September. Comparison of response surface and kriging models for multidisciplinary design optimization. In 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization (p. 4755).

[14] Simpson, T. W., Booker, A. J., Ghosh, D., 2004, "Approximation Methods in Multidisciplinary Analysis and Optimization: A Panel Discussion," Structural and Multidisciplinary Optimization, 27(5) pp. 302-313.

[15] Shan, S., and Wang, G. G., 2010, "Survey of Modeling and Optimization Strategies to Solve High-Dimensional Design Problems with Computationally-Expensive Black-Box Functions," Structural and Multidisciplinary Optimization, 41(2) pp. 219-241.

[16] Cressie, N., 2015, "Statistics for spatial data," John Wiley & Sons, .

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2018, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

[17] Sacks, J., Welch, W. J., Mitchell, T. J., 1989, "Design and Analysis of Computer Experiments," Statistical Science, pp. 409-423.
[18] Bohlin, T.P., 2006, "Practical grey-box process identification: theory and applications," Springer Science & Business Media, .
[19] Psichogios, D. C., and Ungar, L. H., 1992, "A Hybrid Neural Network-first Principles Approach to Process Modeling," AIChE Journal, 38(10) pp. 1499-1511.
[20] Schubert, J., Simutis, R., Dors, M., 1994, "Hybrid Modelling of Yeast Production processes–Combination of a Priori Knowledge on Different Levels of Sophistication," Chemical Engineering & Technology, 17(1) pp. 10-20.
[21] Duarte, B. P., and Saraiva, P. M., 2003, "Hybrid Models Combining Mechanistic Models with Adaptive Regression Splines and Local Stepwise Regression," Industrial & Engineering Chemistry Research, 42(1) pp. 99-107.
[22] Thibault, J., Van Breusegem, V., and Chéruy, A., 1990, "On-line Prediction of Fermentation Variables using Neural Networks," Biotechnology and Bioengineering, 36(10) pp. 1041-1048.
[23] Gibson, I., Rosen, D.W., and Stucker, B., 2010, "Additive manufacturing technologies," Springer, .
[24] Melchels, F. P., Domingos, M. A., Klein, T. J., 2012, "Additive Manufacturing of Tissues and Organs," Progress in Polymer Science, 37(8) pp. 1079-1104.
[25] Bourell, D. L., Beaman, J., Leu, M. C., 2009, "A Brief History of Additive Manufacturing and the 2009 Roadmap for Additive Manufacturing: Looking Back and Looking Ahead," Proceedings of RapidTech, pp. 24-25.
[26] Murr, L. E., Gaytan, S. M., Ramirez, D. A., 2012, "Metal Fabrication by Additive Manufacturing using Laser and Electron Beam Melting Technologies," Journal of Materials Science & Technology, 28(1) pp. 1-14.
[27] Witherell, P., Feng, S., Simpson, T. W., 2014, "Toward Metamodels for Composable and Reusable Additive Manufacturing Process Models," Journal of Manufacturing Science and Engineering, 136(6) pp. 061025.
[28] Agarwala, M., Bourell, D., Beaman, J., 1995, "Direct Selective Laser Sintering of Metals," Rapid Prototyping Journal, 1(1) pp. 26-36.
[29] Cressie, N., 1993, "Statistics for Spatial Data: Wiley Series in Probability and Statistics," Wiley-Interscience, New York, 15pp. 105-209.
[30] Martin, J. D., 2009, "Computational Improvements to Estimating Kriging Metamodel Parameters," Journal of Mechanical Design, 131(8) pp. 084501.
[31] McKay, M. D., Beckman, R. J., and Conover, W. J., 2000, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," Technometrics, 42(1) pp. 55-61.
[32] Louvis, E., Fox, P., and Sutcliffe, C. J., 2011, "Selective Laser Melting of Aluminium Components," Journal of Materials Processing Technology, 211(2) pp. 275-284.

Denno, Peter; Lu, Yan; Witherell, Paul; Krishnamurty, Sundar; Grosse, Ian; Eddy, Douglas.
"investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing."
Paper presented at ASME 2017 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2018, Cleveland, OH, United States. August 6, 2017 - August 9, 2017.

Production System Identification with Genetic Programming

Peter Denno[a,1], Charles Dickerson[b] and Jenny Harding[b]
[a]National Institute of Standards and Technology, US
[b]Loughborough University

Abstract. Modern system-identification methodologies use artificial neural nets, integer linear programming, genetic algorithms, and swarm intelligence to discover system models. Pairing genetic programming, a variation of genetic algorithms, with Petri nets seems to offer an attractive, alternative means to discover system behaviour and structure. Yet to date, very little work has examined this pairing of technologies. Petri nets provide a grey-box model of the system, which is useful for verifying system behaviour and interpreting the meaning of operational data. Genetic programming promises a simple yet robust tool to search the space of candidate systems. Genetic programming is inherently highly parallel. This paper describes early experiences with genetic programming of Petri nets to discover the best interpretation of operational data. The systems studied are serial production lines with buffers.

Keywords. System identification, Petri nets, genetic programming, smart manufacturing

1. Introduction

The ability to generate models of manufacturing systems from data is becoming increasingly useful. In earlier generations of manufacturing systems, a model of the system as a discrete-event system could be developed through inspection of the system's controller software. But, the machine-learning technology that is playing an increasing role in manufacturing control today [1] does not provide a similar presentation of system structure and mechanism. For this reason, verifying system behaviour and safety are becoming more, not less, challenging. Because of this, systems identification methodologies that suggest system structure (grey box models) are becoming increasingly useful. [2]

This paper presents a mostly-automated methodology to produce Petri net (PN) models of systems from historical, operational data describing system inputs and outputs. PNs used in this capacity provide grey-box, system models of the system that suggest the system's structure. Such structurally accurate PN models traditionally have been used for several purposes including verifying system safety, assessing system performance, and detecting deadlocks. This paper describes a method to match differing interpretations of system characterization with the appropriate Petri net structure. For example, multiple interpretations of the notions of blocking and starvation are prevalent in productions-systems engineering. Each interpretation associates accurately with some

[1] Corresponding author

Denno, Peter; Dickerson, Charles; Harding, Jenny.
"Production System Identification with Genetic Programming."
Paper presented at 15th International Conference on Manufacturing Research, London, United Kingdom. September 5, 2017 - September 7, 2017.

Petri net structure. The goal of this work is to determine which interpretation most closely matches the given operational data.

The paper describes preliminary work towards that goal. It uses a case study involving serial production lines. [3] The system-identification algorithm used in this study is provided with two types of inputs: product mixes and machine capacities. The algorithm then generates the corresponding outputs in the form of four, steady-state, performance measures: buffer occupancy, probability of blocking, starvation and throughput. The algorithm applies genetic programming (GP), an evolutionary programming technique, to generalized stochastic Petri nets (GSPNs) to discover a Petri net that best fits the system input-output relations.

The contribution of this paper is a report of early experience using GP with GSPNs. The authors are aware of only one paper of a similar nature [4] and the case study of that paper concerns biological processes, not manufacturing processes, and discrete Petri nets, not timed Petri nets.

Section 2 of the paper discusses related work and the fundamental concepts of Petri nets and genetic programming. Section 3 describes our methodology. Section 4 concludes the work.

2. Related Work

The semantics of GSPNs are described with the help of Figure 1, which depicts a GSPN modelling a system consisting of two machines and a buffer of capacity one between them. The figure, as is typical of Petri net notation, uses 1) solid bars to represent immediate transitions and 2) hollow bars to represent exponentially timed transitions [5] 3) hollow circles to represent places, which can be populated with tokens, and 4) solid circles to represent *m1-blocked*, *buffer* and *m2-busy*. States of the system are described by the quantities of tokens in places. Figure 1 depicts a GSPN describing a system consisting of two machines that process jobs using times from exponential distributions. Between the first machine and the second is a buffer with capacity for one job. In the net on the left side of the figure, both machines *m1* and *m2* are busy; on the right side, machine *m1* is blocked.

GSPNs allow two kinds of arcs between transitions and places, both used in the figure. Activating arcs have an arrow head. Inhibitor arcs have a hollow circle head. A multiplicity is associated with arcs. The execution semantics of GSPNs is as follows. A transition fires when 1) all the input places (places at the tail of an arc) have quantities of tokens equal to or greater than the multiplicity of the arc into the transition, and no input places have quantities of tokens equal to or exceeding the multiplicity of an inhibitor arc into the transition. When the transition fires, one or more tokens, equal to the multiplicity of the input activator arcs, are removed from the input places. Token are added to the output places based on the multiplicity of the output activator arcs. The multiplicity of all arcs in the figure is 1.

Petri nets are one among several representations applied in system identification. Fu and Li [2] survey modern methods of system identification including neural nets, fuzzy logic, genetic algorithms, and swarm intelligence. Though the authors do not specifically discuss genetic programming, many of their comments regarding genetic algorithms may also apply to genetic programming. With respect using genetic algorithms, they note two benefits: quick convergence and path independence; and, one potential drawback: premature convergence to local optima.

Denno, Peter; Dickerson, Charles; Harding, Jenny.
"Production System Identification with Genetic Programming."
Paper presented at 15th International Conference on Manufacturing Research, London, United Kingdom. September 5, 2017 - September 7, 2017.

Nobile et al. [4] describe a methodology for evolving Petri nets for purposes such as system identification. Their test case is a metabolic process, not manufacturing. In their methodology, places and transitions are partitioned into visible and hidden subsets. Each visible place and transition is permanently associated with a domain quantity. Hidden transitions and places can be subject to removal by the evolutionary operations of crossover and mutation. Nobile et al. do not specifically address timed Petri nets, and therefore, do not suggest a means of setting transition rates. The use of GSPN in the present work necessitates a different set of genetic operators than those described by Nobile et al.

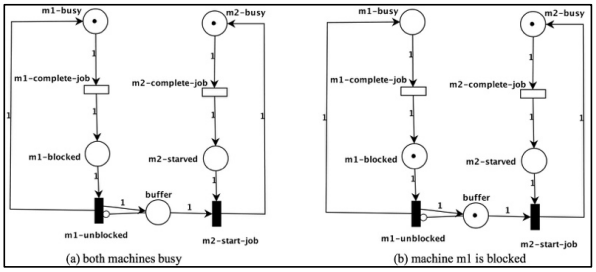


Figure 1. Two machines with a buffer for one part, block-after-service buffering convention.

Several works use Petri nets or genetic algorithms for systems identification. Tiacci [6] couples a discrete event simulator with a genetic algorithm approach to solve an assembly line balancing problem. Dotoli et al. [7] describe a method of system identification using Petri nets and integer linear programming. The work concerns the identification of discrete event systems as untimed Petri nets. In that work, the process is viewed as on-line, in the sense that it waits for events to occur and updates the Petri net after these occurrences. Basile et al. [8] also apply mixed integer linear programming. The Petri nets of this work are deterministic timed, not stochastic. The goal here is to provide a model that matches behaviour as discrete events. Rozinat [9] et al. uses process mining of event logs to create a simulation model of business processes as a coloured Petri Net. The work takes a broad perspective, involving identification of roles and merging of perspectives. Cabasino [10], a PhD thesis, describes an integer programming method of system identification and fault detection using unlabelled Petri nets. The goal of El Medhi et al. [11] is closest to the goal of the present paper. El Medhi describes an identification process for deterministic and stochastic (exponential) Petri net. Such nets can accurately represent queueing systems and machine reliability. The paper describes an integer linear programming method to synthesize a PN from measureable and non-measurable PN states.

3. Genetic Programming of Stochastic Petri Nets

Genetic algorithms are evolutionary algorithms. In a genetic algorithm, a population (sometimes called a generation) of individual solutions are scored for fitness relative to some objective. Those individuals scoring well are more likely to be promoted to the subsequent generation. Those solutions not selected are discarded. Among the promoted individuals, some are subject to modification by the application of two genetic operators:

mutation and crossover. The mutation operator modifies a single, selected individual; The crossover operator swaps elements of two individuals.

Genetic programming [12] is a form of genetic algorithm in which the individuals describe programs, typically represented as trees, and the operators are algebraic. The fitness function scores the ability of such a program to match the input/output relationships provided by training data. In applications where the best match can be represented as a mathematical function, $f: \mathcal{R} \rightarrow \mathcal{R}$, the problem closely resembles regression analysis. Indeed, the problem is a method of system identification called *symbolic regression* (SR). [13] Moreover, it can be viewed as a grey-box method if the discovered system provides a structure corresponding to a physical reality.

Our "programs" are Petri nets. [4] The elements of the programs are, of course, composed of the elements of whatever kind of PN has been selected. In this paper, we are using GSPNs because of their ability to model manufacturing systems. The manufacturing system we intend to examine in this paper concerns a 2-machine, serial production system with exponential service times and a one-place buffer between the machines. System identification in this context involves finding a Petri net that best matches the input/output relationships of the intended system. The chosen outputs are three steady-state properties: buffer occupancy, blocking of machine m1, and starvation of machine m2. These properties were used in the comparing the predicted values against the values of the intended system.

The design space of GSPNs has discrete and continuous dimensions. The discrete dimension concerns the PN's network topology. The continuous dimension concerns the real-valued rates of timed transitions and real-valued weights of immediate transitions. As is typical of generative design problems like SR, some strategy is needed to cope with the discrete/continuous dichotomy. In the design of other SR systems, that strategy uses 1) genetic programming to specify the discrete terms and 2) linear least-squares fitting to determine the optimal values for the continuous elements. Those values are the optimal coefficients of the terms in a linear function that minimizes prediction error. [13]

We have implemented a similar strategy in the design the SR system for our manufacturing example. There, the continuous terms, which are the service times and transition rates, are defined to be exponential with a mean of 1. The discrete term, the topology, evolves through the genetic program. The mutation operators used in that program are responsible for producing variations in the population.

Similar to the one in Nobile et al., [4] our GSPN design method makes a distinction between visible, and hidden, places and transitions. A *visible* place or transition is one for which system observations are provided. For this reason, the visible elements must appear in every PN. A *hidden* place or transition is one not directly associated with a system observation. Therefore, hidden elements need not be included in the PN. Given the visible/hidden dichotomy, the design of our GSPN focuses on visible elements only.

The mutation operators, as well as their arguments, in our GSPN are described in Table 1. In the application of the operators, a modified individual is promoted to the next generation only if it is found to be feasible. (Its reachability graph is calculated to determine this.) If the individual resulting from the mutation fails the feasibility test, a new set of random elements is selected and the individual is retested. If the selected mutation is not possible anywhere in the individual's structure, the individual is not promoted as a mutated form.

Table 1. Mutation operations used in the case study.

Operator	Action

add_place (t1, t2)	Two random transitions, t1, t2 (timed or immediate) are selected and a hidden place, p, and two arcs a1, and a2, are created. a1 is directed from t1 to p. a2 is directed from p to t2. The new arcs have multiplicity 1.
add_token(p)	A token is added to randomly selected place p (visible or hidden).
add_trans(p1, p2)	Two distinct places, p1 and p2 (visible or hidden) are chosen and a timed transition, t, and two arcs, a1 and a2 are created. a1 is directed from p1 to t. a2 is directed from t to p2. The new transition has rate=1.0.
add_arc(p,t)	A random place and transition (without regard to hidden/visible) are selected and randomly, either arc a is directed from p to t or t to p. The multiplicity of the arc is 1.
add_inhibitor(p,t)	A random place, p and transition, t is selected and an inhibitor arc, i, is created and directed from p to t.
remove_place(p)	A random hidden place is selected. It and all arcs to and from it are removed.
remove_token(p)	A place containing at least one token is randomly selected and its token count reduced by 1.
remove_trans(t)	A random hidden transition, t, is selected. It and all arcs to and from it are removed.
remove_arc(a)	A random arc, a, is removed.
remove_inhibitor(i)	A random inhibitor arc, i, is removed.
swap_places(v1, v2)	Two visible places are selected randomly. v1 is assigned the in-coming and out-going arcs of v2 and vice versa.

All selection actions used in the algorithm use tournament selection. Tournament selection involves choosing n individuals randomly from the population and then selecting the best among those n. Thus, when n is high, weak (low scoring) individuals are less likely to be selected. Selection is based on scoring individuals, which involves calculating the steady-state properties of the Petri net and comparing them to the target data. Since the individuals are stochastic Petri nets, this involves 1) forming the infinitesimal generator matrix **Q** for the Markov chain isomorphic to the Petri net and 2) solving the linear system:

$$\begin{aligned} \boldsymbol{\eta Q} &= \mathbf{0} \\ \boldsymbol{\eta 1}^T &= 1 \end{aligned} \tag{1}$$

where $\boldsymbol{\eta}$ is the steady-state distribution vector for the states of the Petri net. [5] For the case study, the dimension of **Q** was typically around 50 for most individuals but ranged as high as 700.

An initial population is created by producing a ring topology-PN called the Eden individual. The Eden individual contains all the visible places and transitions plus additional hidden places or transitions as needed to ensure that there are as many places as transitions. With equal numbers of places and transitions, the ring topology PN is produced by adding arcs between alternating places and transitions and closing the graph by connecting the last element used to the first. The Eden individual is repeatedly subjected to the genetic operators to create all of the individuals in the initial population.

In the case study, the visible places were labelled *m1-busy*, *m2-busy*, *buffer*, *m1-blocked,* and *m2-starved* as depicted in Figure 1. The system under study follows the usual conventions for analysis of serial production systems [3]: the first machine cannot starve and the last machine cannot block. The training data corresponded to machines with exponential service time. ($\lambda = 1.0$ for both machines.) The case study system uses block-after-service blocking convention.

Experience with the case study problem suggests that the algorithm essentially works, but that more effort will be needed to avoid convergence to local optima. As Table 2 shows, the algorithm tended to find a local optimum quickly and stick with it while the

median individual slowly improved. This was the case even when tournament selection pressure was low. The population size of the case study is 100 individuals. Bloat (the tendency in GP for individuals to become increasingly complex with little performance gain) was not a problem.

Table 2: Preliminary results from the case study

Generation	Error of Best Individual (total absolute error)	Error of Median Individual (total absolute error)
0	0.684	> 100.
1	0.684	1.80
2	0.494	1.33
8	0.470	0.855
9	0.333	0.720
10	0.333	0.512
12	0.333	0.467

4. Conclusion

The paper described early experience with a system-identification methodology that applies genetic programming to Petri nets representing discrete-event systems. The goal of the work is to discover Petri nets that best interpret the operational data. Genetic programming promises an effective method that easily parallelizes this problem. Early results suggest that the methodology is sound but that more work will be required to make it effective. We are currently undertaking that work.

References

[1] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine learning in manufacturing: advantages, challenges, and applications," *Prod. Manuf. Res.*, 2017.

[2] L. Fu and P. Li, "The Research Survey of System Identification Method," in *2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics*, 2013, pp. 397–401.

[3] J. Li and S. M. Meerkov, *Production System Engineering*. Springer Science+Business Media, 2009.

[4] M. S. Nobile, D. Besozzi, P. Cazzaniga, and G. Mauri, "The foundation of Evolutionary Petri nets," *CEUR Workshop Proc.*, vol. 988, pp. 60–74, 2013.

[5] M. A. Marsan, G. Balbo, G. Conte, and G. Franceschinis, "Modelling with generalised stochastic petri nets," *System*, p. 299, 1994.

[6] L. Tiacci, "Coupling a genetic algorithm approach and a discrete event simulator to design mixed-model un-paced assembly lines with parallel workstations and stochastic task times," *Int. J. Prod. Econ.*, vol. 159, pp. 319–333, 2015.

[7] M. Dotoli, M. P. Fanti, and A. M. Mangini, "Real Time Identification of Discrete Event Systems by Petri Nets," in *IFAC 2007*, 2007.

[8] F. Basile, S. Member, P. Chiacchio, S. Member, and J. Coppola, "Identification of Time Petri Net Models," *IEEE Trans. Syst. Man Cybern. Syst.*, pp. 1–15, 2016.

[9] A. Rozinat, R. S. Mans, M. Song, and W. M. P. van der Aalst, "Discovering simulation models," *Inf. Syst.*, vol. 34, no. 3, pp. 305–327, 2009.

[10] M. P. Cabasino, "Fault diagnosis and identification of discrete event systems using Petri nets," University of Cagliari, 2008.

[11] S. Ould El Mehdi, R. Bekrar, N. Messai, E. Leclercq, D. Lefebvre, and B. Riera, "Design and Identification of Stochastic and Deterministic Stochastic Petri Nets," *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 42, no. 4, pp. 931–946, 2012.

[12] R. Poli, W. B. Langdon, N. F. Mcphee, and J. R. Koza, "A Field Guide to Genetic Programming," 2008.

[13] D. P. Searson, "GPTIPS 2: an open-source software platform for symbolic data mining," 2015.

Denno, Peter; Dickerson, Charles; Harding, Jenny.
"Production System Identification with Genetic Programming."
Paper presented at 15th International Conference on Manufacturing Research, London, United Kingdom. September 5, 2017 - September 7, 2017.

Quantifying Thermophoretic Deposition of Soot on Surfaces

Amy Mensch, Thomas Cleary
National Institute of Standards and Technology, Gaithersburg, MD, USA

Abstract

Quantitative data on deposition of soot agglomerate particles in the literature is needed to advance fire forensic analysis as well as fire model predictions of visibility and detector activation. This paper provides direct measurements of thermophoretic soot deposition in a laminar flow channel and the driving conditions to improve understanding of soot deposition in fires and for deposition model assessment. The overall deposition velocities were determined through measurements of the incoming soot concentration and gravimetric measurements of the soot deposited. The effects of channel flowrate and temperature gradient as well as inlet concentration were examined. The deposition velocities showed good agreement with the theoretical thermophoretic velocities based on the channel temperature gradients. The flow, heat transfer and deposition were also modeled using the Fire Dynamics Simulator, and the simulation deposition velocities were generally less than those found in the experiments.

Keywords: soot, thermophoresis, deposition velocity

Introduction

The physics of soot deposition in fires are controlled by thermophoretic, turbulent and gravitational deposition mechanisms. The thermophoretic mechanism is driven by temperature gradients in the gas, which impart unequal collision energies on aerosol particles between the hot and cold sides, resulting in motion in the opposite direction of the temperature gradient. The focus of this study is on thermophoretic deposition, which has a significant role in fires, especially for small particles (less than 1 μm) [1] produced during flaming combustion.

Thermophoretic deposition is characterized by the terminal velocity of particles driven by thermophoresis. The theoretical thermophoretic velocity, v_{th}, is proportional to the temperature gradient, ∇T, and the

Mensch, Amy; Cleary, Thomas.
"Quantifying Thermophoretic Deposition of Soot on Surfaces."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

kinematic viscosity, ν, of the gas, and inversely related to the temperature of the particle, T_p:

$$v_{th} = -K_{th} \frac{\nu \nabla T}{T_p} \quad \text{(Eq. 1)}$$

K_{th} is the thermophoretic coefficient. For a Knudsen number (the ratio of gas mean free path to particle radius), Kn >> 1, when the mean free path of the gas is much greater than the particle size, K_{th} is generally assumed to be 0.55 and independent of particle size [2]. This condition is known as the free molecular regime. K_{th} can also be calculated as a function of Kn (which is a function of particle size and gas temperature), the thermal conductivities of the gas and particle, and empirical constants [3]. Studies of soot agglomerates have generally found that K_{th} should be evaluated using the primary particle diameter [4]–[6], suggesting the use of K_{th} = 0.55 for soot. Suzuki et al. [6] noted that K_{th} also depends on the morphological characteristics of the agglomerates, with more open structures being closer to the free molecular regime compared with more compact agglomerates.

Currently, there is insufficient validation data to assess the performance of predictive models of thermophoretic soot deposition. Researchers have used different approaches to measure soot deposition from fires, including optical scanning of glass paper deposition targets [7], directly measuring physical thickness [8], and measuring the response of a conductometric gauge [9]. Several soot deposition studies [4], [10], [11], motivated by the need to monitor soot in diesel exhaust, generated either indirect or qualitative measurements of surface deposition.

Soot deposition experiments

The mechanism behind thermophoretic deposition was studied within a thin rectangular laminar flow channel with a transverse temperature gradient applied across the channel height, ensuring that deposition occurred only on the cold side of the channel. The channel was positioned so the flow was vertically downward to remove the effect of gravitational deposition. The flow passed through a plenum before and after the flow channel to minimize entrance and exit effects. The side walls were polytetrafluoroethylene, and the cold and hot boundaries were aluminum slabs, 19.1 mm thick to approximate constant temperature boundaries. On the outer side of the cold wall was a serpentine copper line circulating cold water, resulting in outer cold wall temperatures between 15 °C and 20 °C. On the outer side of the hot wall was a resistance heater with on-off control based on a set point temperature measured on the outer side of the hot wall. This temperature was set to 230 °C and 120 °C to generate cases with internal temperature differences of approximately 200 °C and 100 °C.

A steady-state flow simulation, which included the inlet plenum geometries, predicted fully-developed flow and temperature profiles by

Mensch, Amy; Cleary, Thomas.
"Quantifying Thermophoretic Deposition of Soot on Surfaces."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

20 % of the channel length for a channel flow of 3 SLM (standard L/min) and 50% of the channel length for 10 SLM [12]. The flow was laminar for both flowrates, using the Reynolds number based on the hydraulic diameter of the channel (< 230 for all cases). The fully-developed temperature profiles were confirmed to be linear across the channel height. The channel geometry is depicted in Fig. 1.

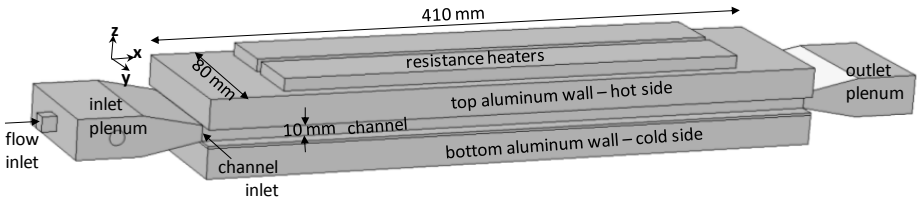


Fig. 1 Diagram of experimental channel geometry

A laminar diffusion flame burner was used to generate soot for deposition. Propene fuel exited a 10 mm diameter tube surrounded by co-flow air from a 120 mm diameter ceramic honeycomb, enclosed by a brass chimney. After a tripper plate to induce mixing, additional dilution air was injected into the upper stage of the burner. All fuel and air flows to the burner, 0.055 SLM and 0.077 SLM for the fuel, 54.08 SLM for the co-flow air, and 32.47 SLM for the dilution air, were set by mass flow controllers. The duration of deposition exposures ranged from 15 min to 60 min. The soot aerosol concentration, C_p, entering the channel was measured by two methods, by flowing part of the exhaust from the burner through a tapered element oscillating microbalance (TEOM), and by flowing part of the exhaust through a filter to measure the change in mass captured at the measured flowrate. The averages and expanded uncertainties at 95 % confidence interval (μ) of C_p are given in Table 1.

Table 1. Particle concentration, C_p, measurements

Fuel flow (L/min)	$C_{p,ave}$ (mg/m^3)		No. of experiments	μ of $C_{p,ave}$ (mg/m^3)	μ of C_p, in each experiment (mg/m^3)
0.055	TEOM	**66**	21	± 21.0	± 6.1
	Filter	**70**	15	± 5.4	± 7.5
0.077	TEOM	**108**	3	± 10.4	± 39.0
	Filter	**125**	11	± 19.2	± 11.4

At the end of the exposure the mass loading of soot deposited on the cold side of the channel was determined gravimetrically by measuring the change in mass, m_{dep}, on four aluminum foil circular targets (each with A_{dep} = 1.7E-03 m^2). The targets were spaced along the channel centerline, centered at 46 mm, 148 mm, 249 mm, and 351 mm from the channel inlet. The mass of the targets was taken at least several hours

Mensch, Amy; Cleary, Thomas.
"Quantifying Thermophoretic Deposition of Soot on Surfaces."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

after the end of an experiment, after the channel was cooled to room temperature. Subsequent mass measurements performed after desiccating overnight did not show consistent reductions to indicate condensed volatiles or water on the targets. The uncertainty in mass loading measurements was estimated as ± 1 mg/m^2. The standard deviation over the average of the four mass loading measurements was 15 % or less for all but one experiment, with the biggest deviation coming from the first target. The average mass loading, the incoming C_p, and the exposure time, t, were used in Eq. 2 to calculate the overall deposition velocity, v_{dep}. The filter-based C_p was used for experiments when the filter was included. Otherwise, the TEOM-based C_p was used.

$$v_{dep} = \frac{m_{dep}}{C_p A_{dep} t} \qquad \text{(Eq. 2)}$$

The average v_{dep} from each experiment could be compared to the theoretical v_{th} determined from Eq. 1 because the thermophoretic force was the major mechanism driving particles to deposit, and the flow and C_p were steady throughout the experiments. Because of the linear temperature gradient across the channel, the ∇T for the calculation of v_{th} was determined based on surface temperature measurements from steady-state experiments without deposition [12]. The average of the temperatures measured at the inlet and outlet was used to estimate T_p. Table 2 reports ∇T and T_p for the different cases of flow and ΔT. A K_{th} of 0.36 was used to achieve the best matching of v_{th} with v_{dep}.

Table 2. Channel flow and thermal measurements

	ΔT = 200 °C		ΔT = 100 °C	
Channel flow (SLM)	3.00	10.00	3.00	10.00
Measured ∇T (°C/m)	19745	19704	10394	10408
Measured T_p (°C)	69	77	47	57

The results comparing the measured v_{dep} and the calculated v_{th} are plotted in Fig. 3 with the dashed line representing correspondence between the two velocities. Each symbol represents one experiment, with different color symbols representing different fuel flows or measurement methods for C_p. The error bars show the estimated combined expanded uncertainties, which were ±15 % for v_{th} and varied from ±7.5 % to ±11.3 % for v_{dep}. The primary factor that affected deposition velocity was the applied temperature gradient of the exposure. The cases with ΔT of 200 °C are clustered between 0.4 mm/s and 0.5 mm/s, and the cases with ΔT of 100 °C are clustered around 0.2 mm/s. One data point had error bars outside of the dotted line (v_{dep} = 0.14 mm/s, v_{th} = 0.21 mm/s). This point was for a channel flow of 10 SLM and ΔT of 100 °C. For this case the mass deposited on the first target was significantly lower than the average of the four targets, and its standard deviation over the average was 24 %. The local

temperature gradient close to the cold surface could be lower than expected before the thermal profile was fully-developed, and therefore have locally reduced deposition. Indirectly, the channel flow could affect deposition velocity through changes to the flow and temperature profiles. However, the cases with 10 SLM and ΔT of 200 °C did not have significantly less deposition on the first target, and the overall average deposition velocities were close to the dotted line in Fig. 3. The C_p and fuel flow also could have indirectly affected deposition velocity through changes to the soot size distribution, but there was no apparent distinction in cases with different fuel flowrates. Therefore, it was found that temperature gradient, and not C_p, fuel flow, or channel flow, directly affected deposition velocity, as expected based on Eq. 1.

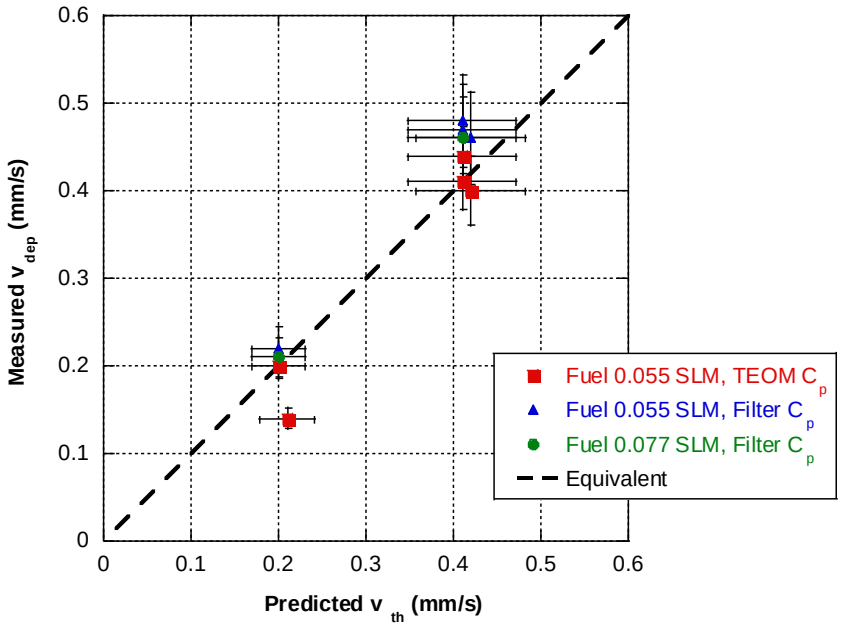


Fig. 3 Experimental v_{dep} versus predicted v_{th}

Computational modelling

To model the soot deposition within the channel, computational flow and heat transfer simulations were run using the NIST Fire Dynamics Simulator (FDS) [3], [13]. The simulations were transient to track the buildup of soot on the surface. In FDS's aerosol deposition models, soot was treated as an additional gaseous species for which diffusive transport along concentration gradients was calculated automatically. To account for thermophoretic transport, an additional velocity based on Eq. 1 was applied to the aerosol species. Soot was introduced into the inlet flow at a concentration of 70 mg/m^3, although this value did not affect v_{th} or the distribution of soot deposition, only the amount of deposit. The default soot properties from FDS were used, except for particle diameter, which was specified to be 0.035 µm, as an estimate

of the primary particle size of soot agglomerates [6]. FDS used the temperature in the first grid cell above the wall for T_p and the temperature dependent properties needed to calculate v_{th}. The K_{th} was calculated to be 0.55 in the FDS simulations.

The computational mesh was a structured rectangular grid with spacing of 5 mm across the width and length, and 1 mm across the height. The simulations were run for 1000 s, with steady deposition rates reached by 10 s. The hot and cold wall boundary conditions were constant temperature, and the side walls were adiabatic to approximate the experimental conditions. The inlet flow profile was prescribed based on the steady-state flow solution that included the inlet plenum geometry, while the flow temperature for the inlet was determined from the measurements within the inlet plenum.

Fig. 4 shows the FDS results for deposition velocity, v_{FDS}, just above the cold wall. The value of v_{FDS} covering the largest area in the downstream portion of the channel is labeled for each case. When ΔT was doubled from 100 °C to 200 °C, the v_{FDS} results were slightly more than doubled. When the channel flow increased from 3 SLM to 10 SLM, v_{FDS} decreased slightly for both temperature differences. The v_{FDS} increased along the flow direction as the flow and temperature profiles developed. The increases in v_{FDS} continued farther into the channel for the 10 SLM cases compared to the 3 SLM cases, which were more stable in the downstream half of the channel. These differences between the channel flowrates were attributed to the differences in the flow development, with 10 SLM requiring a longer channel distance to completely develop. In general, v_{FDS} predictions were lower than v_{dep} and v_{th} from corresponding experiments, except for the downstream v_{FDS} for the 3 SLM cases, which were close to the v_{dep} and v_{th}.

Two significant aspects of the FDS thermal predictions are affecting the deposition velocity comparisons with the experiments. First, the development length for fully-developed temperature profiles in FDS is longer than expected in the experiments based on the detailed steady-state flow simulations [12]. Fully-developed profiles in the experiments are also confirmed by the uniformity of soot loading measurements across the length, particularly for the final three out of four targets. The second discrepancy is with the values calculated for ∇T in FDS, which are significantly less than the measured ∇T's in Table 2. FDS determines ∇T using wall heat transfer coefficient correlations [3], which may be causing errors to v_{FDS} for these cases of laminar channel flow.

Conclusion

Laminar flow through a thin rectangular channel with a transverse temperature gradient was used to generate thermophoretic deposition exposures on a target surface for different cases of channel flowrate, channel temperature gradient, and fuel flowrate. The mass of deposition

Mensch, Amy; Cleary, Thomas.
"Quantifying Thermophoretic Deposition of Soot on Surfaces."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

was measured gravimetrically and combined with measurements of the inlet soot concentration to determine the overall deposition velocity. The deposition velocity compared well with the predicted thermophoretic velocities based on the channel temperature gradient and an assumed thermophoretic coefficient. The channel flow was also modeled with FDS to generate predictions of soot deposition in the channel. The simulated deposition velocities showed the expected trends with temperature gradient, but were generally lower than the experimental deposition and thermophoretic velocities. The differences in the deposition velocities and thermophoretic coefficients were attributed to FDS predicting slower thermal development compared to the experiments, and to FDS calculating the temperature gradient from the heat transfer coefficient, rather than from the overall channel temperature difference.

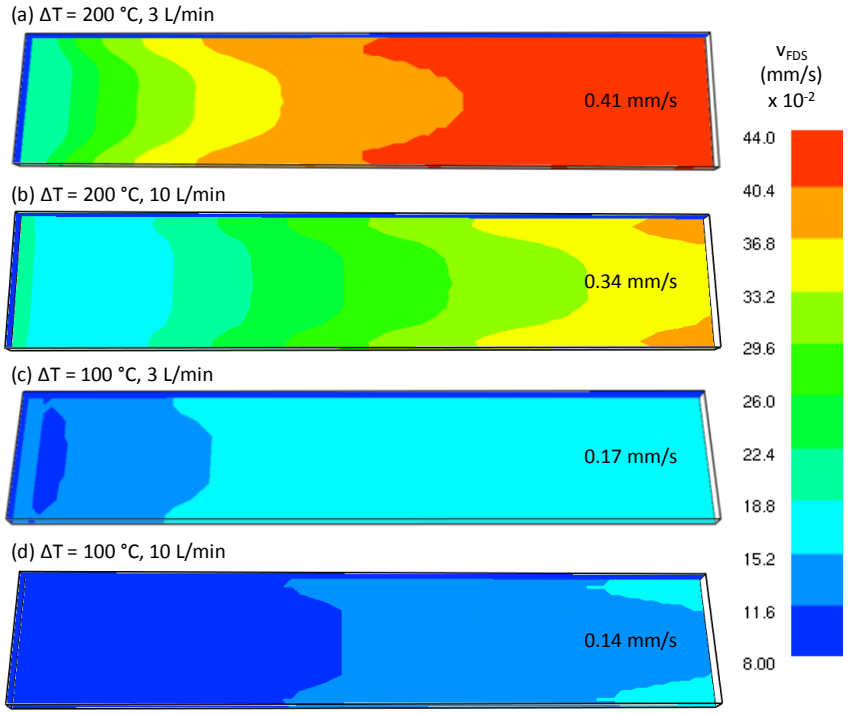


Fig. 4 FDS results for deposition velocity distribution

References

[1] K. M. Butler and G. W. Mulholland, “Generation and Transport of Smoke Components,” *Fire Tech.*, vol. 40, no. 2, pp. 149–176, Apr. 2004.

[2] W. C. Hinds, *Aerosol technology: properties, behavior, and measurement of airborne particles*, 2nd ed. New York: Wiley, 1999.

[3] K. McGrattan, S. Hostikka, R. McDermott, J. Floyd, C. Weinschenk, and K. Overholt, “Fire Dynamics Simulator Technical Reference Guide Volume 1: Mathematical Model,” NIST Special Publication 1018–1, Apr. 2015.

[4] A. Messerer, R. Niessner, and U. Pöschl, “Thermophoretic deposition of soot aerosol particles under experimental conditions relevant for modern diesel engine exhaust gas systems,” *J. of Aerosol Science*, vol. 34, no. 8, pp. 1009–1021, Aug. 2003.

[5] D. E. Rosner and Y. F. Khalil, “Particle Morphology- and Knudsen Transition-Effects on Thermophoretically Dominated Total Mass Deposition Rates From ‘Coagulation-Aged’ Aerosol Population,” *J. of Aerosol Science*, vol. 31, no. 3, pp. 273–292, Mar. 2000.

[6] S. Suzuki, K. Kuwana, and R. Dobashi, “Effect of particle morphology on thermophoretic velocity of aggregated soot particles,” *Int. J. of Heat and Mass Transfer*, vol. 52, no. 21–22, pp. 4695–4700, Oct. 2009.

[7] S. Riahi and C. Beyler, “Measurement and Prediction of Smoke Deposition from a Fire Against a Wall,” *Fire Safety Science*, vol. 10, pp. 641–654, 2011.

[8] W. D. Ciro, E. G. Eddings, and A. F. Sarofim, “Experimental and Numerical Investigation of Transient Soot Buildup on a Cylindrical Container Immersed in a Jet Fuel Pool Fire,” *Comb. Sci. and Tech.*, vol. 178, no. 12, pp. 2199–2218, Jun. 2006.

[9] T. Cleary, “Effects of Soot Deposition on Current Leakage in Electronic Circuitry,” in *Proc. of the 15th Int. Conf. on Automatic Fire Detection (AUBE)*, 2014.

[10] G. Hagen, C. Feistkorn, S. Wiegärtner, A. Heinrich, D. Brüggemann, and R. Moos, “Conductometric Soot Sensor for Automotive Exhausts: Initial Studies,” *Sensors*, vol. 10, no. 3, pp. 1589–1598, Mar. 2010.

[11] D. Lutic *et al.*, “Detection of Soot Using a Resistivity Sensor Device Employing Thermophoretic Particle Deposition,” *J. of Sensors*, vol. 2010, p. 421072, Jun. 2010.

[12] A. Mensch and T. Cleary, “EL Exploratory Project: A Soot Deposition Gauge for Fire Measurements,” National Institute of Standards and Technology, NIST TN, In progress.

[13] K. McGrattan, R. McDermott, C. Weinschenk, K. Overholt, S. Hostikka, and J. Floyd, “Fire Dynamics Simulator User’s Guide 6th Edition,” NIST Special Publication 1019, Apr. 2015.

Mensch, Amy; Cleary, Thomas.
”Quantifying Thermophoretic Deposition of Soot on Surfaces.”
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE ’17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

A Comparison of Carbon Monoxide Gas Sensing to Particle Smoke Detection in Residential Fire Scenarios

Thomas Cleary, Amy Mensch
National Institute of Standards and Technology, Gaithersburg, MD, USA

Abstract

Recent research has suggested that carbon monoxide (CO) sensing might be better than photoelectric detectors for detecting smoldering fires in dwellings. Results from that research were compared to full-scale experimental data sets, where carbon monoxide concentration and smoke alarm response were gathered during smoldering polyurethane foam furniture and furniture mockup experiments. Based on the analysis of those data sets, CO gas sensing is complementary to particulate smoke detection, but does not appear to rise to a level suggesting it should be a required in a standalone smoke detector.

Keywords: Smoke alarms, carbon monoxide detection, smoldering fire

Introduction

Sesseng and Reitan presented research they assert demonstrated that carbon monoxide (CO) sensing might be better than photoelectric detectors for detecting smoldering fires in dwellings [1]. Specifically, they claimed that photoelectric detectors may not be safe in a smoldering fire because a sleeping occupant may be overcome by CO before a photoelectric alarm triggers, and earlier notification of the fire brigade from CO detection may save lives and reduce property damage. While they do acknowledge the need for particulate smoke detection of flaming fires where CO production is relatively low and fire development is rapid, their research poses a question: Should CO sensing be a requirement for residential fire detection?

While research has demonstrated the utility of CO and other gas sensing in early fire detection, there is a lack of analyzed data suggesting its superiority over particulate sensing. To provide guidance in answering the question above, the experimental set-up used by Sesseng and Reitan is described and critiqued as to its relevance in mimicking realistic scenarios, and their results are compared to full-scale experiments conducted by the National Institute of Standards and Technology (NIST).

Cleary, Thomas; Mensch, Amy.
"A Comparison of Carbon Monoxide Gas Sensing to Particle Smoke Detection in Residential Fire Scenarios."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

Sesseng and Reitan Experiments

The experiments were conducted in a test room of interior dimensions 3.6 m × 2.4 m × 2.4 m high which met the ISO 9705 Standard [2]. This room size is typical of a small bedroom with a volume of 21 m^3. The door opening was kept closed during the experiments. A piece of polyurethane foam mattress 0.7 m × 0.5 m × 0.1 m was the fuel source. To initiate smoldering, a resistance heating wire was wrapped in cotton batting and placed on top of the mattress segment. The entire mattress segment was covered with insulating ceramic fiber blanket material. Finally, a wooden box with a 51 cm diameter hole in the center of the top surface was placed over the mattress segment to force the smoke through that central aperture. The heater was energized to about 23 W for 10 minutes to initiate smoldering in the foam.

Ten experiments were conducted with the smoldering source placed in various locations including on a raised platform representing a bedframe, under the platform, and on the floor. Carbon monoxide concentration was measured at a location representing the toxic gas exposure of an occupant sleeping on the bed platform. The CO concentrations measured at that location were consistent with CO sensors placed in various locations of the room, indicating an even distribution of CO in the room. Those concentrations at the minimum and average alarm times for both CO and photoelectric alarms were tabulated. Additionally, a CO dose was computed by integrating the CO concentration (reported as volume fraction×10^6 or ppm) as a function of time, up to the point of alarm. The CO dose was compared to a median incapacitation dose (IC_{50}) of 35,000 ppm×min.

First, it is noted that the room was relatively small and unventilated, which would be a worst-case scenario for CO build-up for a given source. Second, it appears that the ceramic blanket would filter some of the smoke particulate while allowing gaseous CO to diffuse through it and then throughout the room. Third, the confining box, while providing a repeatable location for the smoke to emanate from, most-likely affected the natural plume(s) from the smoldering foam, affecting buoyancy and the transport of smoke to the ceiling.

It seems that the experimental set-up was likely to produce results where CO sensing would outperform particulate sensing and may not mimic realistic smoldering upholstered furniture fire scenarios. Nonetheless, their results taken at face value demand a more detailed evaluation of existing data from smoldering fire experiments to make a judgement if CO sensing should be considered as a requirement.

NIST Experiments

Two full-scale experimental data sets were analyzed to compare carbon monoxide sensing to photoelectric or ionization smoke alarm response

Cleary, Thomas; Mensch, Amy.
"A Comparison of Carbon Monoxide Gas Sensing to Particle Smoke Detection in Residential Fire Scenarios."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

in smoldering furniture or polyurethane foam chair mockup fire scenarios.

The first data set was the NIST smoke alarm sensitivity study where chair mockups, consisting of non-fire-retarded polyurethane foam covered with cotton cushion covers, were smoldered [3]. Figure 1 shows the mockup and the ignition set-up. The cushions rested on a metal frame that was placed on a raised platform attached to a load cell. A small square of cotton fabric was place at a front corner location and a 50 W electric cartridge heater about the size of a cigarette was place on the fabric. After energizing the heater for about 6 minutes, the heater was removed and seat cushions smoldered. Eventually, smoldering reached the back cushion and transitioning to flaming in about 90 minutes on average in 11 of 12 experiments. One chair mockup did not transition to flaming before the end of the experiment.

Figure 1. Chair mockup consisting of polyurethane foam slabs with cotton seat cushion and chair back cushion covers.

Figure 2 is a schematic of the small apartment mockup experimental space. Experiments were conducted with the smoldering source located in the living room with the door to the master bedroom closed as shown, and in the master bedroom with the door either open or closed. Twelve initially smoldering fire experiments were conducted, six with the source in the living room, and three each with the source in the master bedroom with the door open or closed. The volume of the master bedroom was 38 m^3 and the volume of the living room and attached spaces excluding the master bedroom was 92 m^3.

Photoelectric and ionization smoke alarms were located at various ceiling locations. Gas samples were extracted from a height of 1.5 m from the floor at the locations indicated. Here, alarm times for photoelectric and ionization alarms in either the master bedroom (S5 or S6) or hallway locations (S2 or S3) were tabulated along with the CO concentration at alarm from the nearest sampling location and the corresponding computed fractional effective dose of the toxic gases (FED) [4]. For the smoldering phase, the toxic gases considered were CO and hydrogen cyanide. (In these experiments, hydrogen cyanide grab samples were analyzed and correlated to CO concentration during

Cleary, Thomas; Mensch, Amy.
"A Comparison of Carbon Monoxide Gas Sensing to Particle Smoke Detection in Residential Fire Scenarios."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

the smoldering and flaming stages of combustion [5].) The fractional effective dose for toxic gases increased more rapidly when estimated hydrogen cyanide concentration was included.

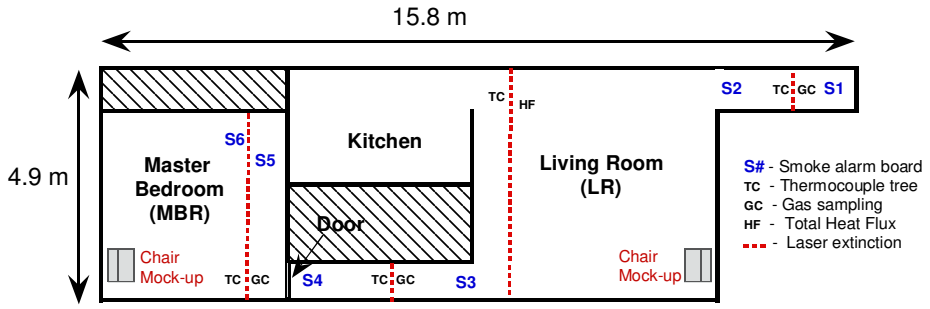


Figure 2. Schematic of the small apartment mockup space showing the location of the smoldering sources.

The second data set is from the NIST home smoke alarm project where upholstered chairs and mattresses were smoldered by inserting an electric wire resistance heater into a slit in the covering fabric and foam of the chairs or mattresses [6]. The experiments were conducted in a single-story manufactured home and a two-story home slated for demolition. The experiments were conducted without any forced ventilation. Individually calibrated smoke and CO alarms were installed at various ceiling locations in groups that included multiple photoelectric, ionization, CO alarms. Alarms were calibrated with smoldering cotton wick smoke in the NIST fire emulator/detector evaluator tunnel. Alarm points were chosen as 6.4 %/m obscuration (2.0 %/ft. in U.S. industry standard units) and 50 ppm for photoelectric and CO alarms respectively. Alarm locations in hallways adjacent to the room of fire origin were selected, and the average time to reach the alarm threshold for the photoelectric and CO alarms was computed.

Comparison of Results

Tabulated results from each of the experimental data sets is presented below. First, Table 1 shows the results from Sesseng and Reitan. In addition to their computed dose, a computed FED for CO was tabulated for each averaged photoelectric and CO alarm time by dividing the CO dose by 35,000 ppm×min to facilitate comparison to other experimental data. A FED of 1.0 indicates an exposure that incapacitates 50% of a normally susceptible population [4].

In four out of seven cases, the averaged photoelectric alarm time yielded a FED greater than 1.0, hence at least half of sleeping occupants exposed may not have been alerted prior to an incapacitating dose. Conversely, in all cases the average CO alarm time yielded very low FED exposures presumably providing alert to all sleeping occupants.

Cleary, Thomas; Mensch, Amy.
"A Comparison of Carbon Monoxide Gas Sensing to Particle Smoke Detection in Residential Fire Scenarios."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

Table 1. Results from Sesseng and Reitan for the CO concentration, dose and FED at the average photoelectric or CO alarm time for each experiment [1].

Exp. #	Photoelectric Alarm			CO Alarm		
	CO (ppm)	CO Dose (ppm×min)	FED	CO (ppm)	CO Dose (ppm×min)	FED
1	664	37593	1.07	35	875	0.03
2	1453	63957	1.83	42	766	0.02
3	638	24371	0.70	62	1236	0.04
4	907	39325	1.12	61	965	0.03
6	933	32547	0.93	35	315	0.01
7	1075	64184	1.83	37	554	0.02
8				46	1019	0.03
9				36	489	0.01
10				46	960	0.03

Table 2 shows the results from the NIST smoke alarm sensitivity study for both photoelectric and ionization alarms. The FED computation includes the effects of hydrogen cyanide, thus is more conservative than computed values of CO alone.

Table 2. Results from the NIST smoke alarm sensitivity study for photoelectric and ionization alarms [3].

Experimental Configuration	Photoelectric Alarm		Ionization Alarm	
	CO (ppm)	FED	CO (ppm)	FED
BR door closed	34	0.02	26	0.01
BR door closed	104	0.05	46	0.02
BR door closed	20	0.01	40	0.01
BR door closed	25	0.01	44	0.01
BR door opened	22	0.01	15	0.01
BR door opened	16	0.01	20	0.01
BR door opened	37	0.01	45	0.02
LR	135	0.03	250	0.03
LR	-	-	260	0.03
LR	85	0.05	375	0.17
LR	50	0.02	400	0.20
LR	88	0.01	202	0.03
LR	-	-	25	0.002
LR	30	0.02	29	0.004
LR	40	0.01	62	0.03
LR	-	-	54	0.02

Cleary, Thomas; Mensch, Amy.
"A Comparison of Carbon Monoxide Gas Sensing to Particle Smoke Detection in Residential Fire Scenarios."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

The concentration of CO at alarm was lower on average for photoelectric alarms than for ionization alarms, and the computed FED was below 0.1 for all average photoelectric alarm times and above 0.1 for only two average ionization alarm times.

Figure 2 is a scatter plot of all FED values from the Sesseng and Reitan experiments (SP) and the NIST smoke alarm sensitivity study. This plot illustrates the difference between Sesseng and Reitan's photoelectric alarm results and their CO alarm and NIST smoke alarm results. The differences in room volume range from 21 m^3 in the SP study to 38 m^3, 92 m^3, and 130 m^3 for the various experimental configurations in the NIST study. While room size may have influenced CO concentration, it was observed in the NIST study that smoke alarms tended to respond much sooner when the source and alarms were confined to the smaller master bedroom space.

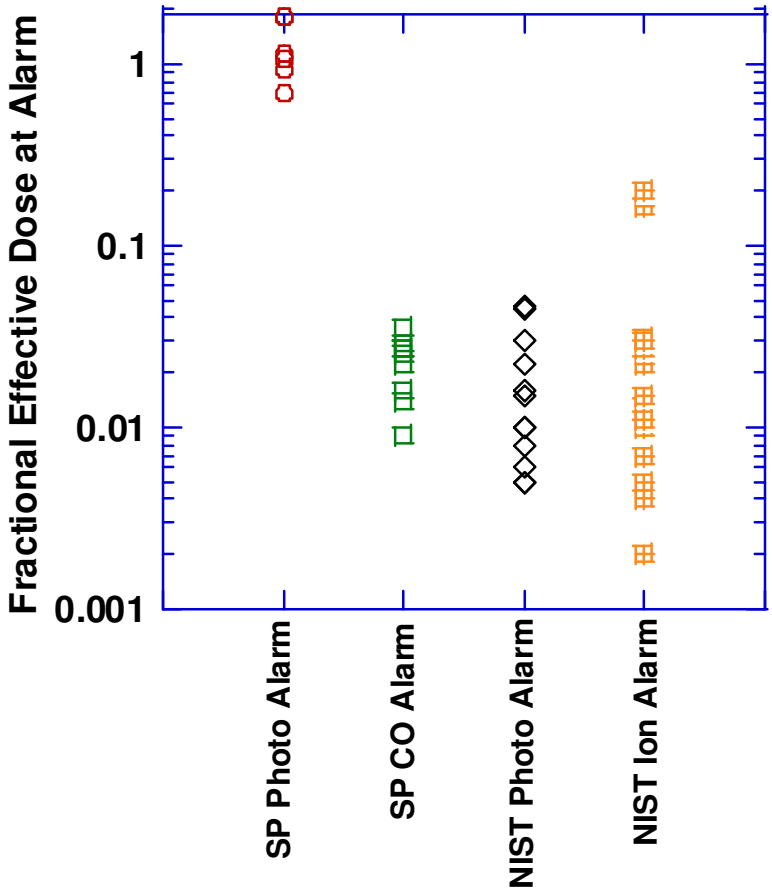


Figure 2. FED values from the Sesseng and Reitan experiments (SP) [1] and the NIST smoke alarm sensitivity study [3] at various average alarm times.

Table 3 shows results from smoldering chairs and mattresses in the NIST home smoke alarm study. Those results compare the average time to alarm for co-located CO alarms (with a calibrated alarm concentration of 50 ppm) and co-located photoelectric alarms (with a smoke box alarm obscuration of 6.4 %/m). Also tabulated are CO concentrations at 1.5 m from the floor at the average photoelectric alarm times.

Table 3. Results from the NIST home smoke alarm study [6].

Experiment	Avg CO Alarm (s)	Avg Photoelectric Alarm (s)	CO conc, at Photoelectric alarm (ppm)
SDC01	3302	5382	230
SDC04	3403	1153	-
SDC06	4741	3473	-
SDC11	3942	4241	117
SDC31	5092	5041	225
SDC34	-	-	100
SDC37	-	-	50
SDC40	-	-	38
SDC23	4599	4664	-
SDC27	2761	1366	-

The average time to CO alarm was shorter in only three of seven experiments. The CO concentration at the average photoelectric alarm time was significantly lower than the values recorded by Sesseng and Reitan.

Conclusions

Analysis of the NIST data sets showed photoelectric detection in the room of fire origin was sufficient in all smoldering fire cases to provide early warning prior to hazardous CO exposures at the specific locations. CO detection may provide significantly earlier warning than ionization alarms for some smoldering scenarios which could provide earlier notification to the fire brigade. However, the new fire test requirements of ANSI/UL 217-2015 [7] will improve alarm response to smoldering upholstered furniture fires containing polyurethane foam, ameliorating the relatively slower response of ionization alarms compared to photoelectric alarms for such smoldering fire scenarios.

Based on the analysis of existing data sets, CO gas sensing can be complementary to particulate smoke detection, but does not appear to

rise to a level suggesting it should be a required in a standalone smoke detector. Nonetheless, the new fire and cooking nuisance tests introduced in ANSI/UL 217-2015 may provide an incentive for smoke alarm designs to include CO gas sensing for nuisance alarm resistance. In addition, several manufacturers currently produce combination smoke / CO alarms combining the functions of standalone smoke and CO alarms. Smoke alarm manufacturers may find benefits in considering CO gas sensors to compliment smoke alarm activation in smoldering fires as a detection enhancement.

Acknowledgements and Disclaimer

Official contribution of the U.S. Government; not subject to copyright in the United States.

References

[1] Sesseng, Christian and Reitan, Nina Kristine, "*Experimental investigation of using CO sensors to detect smouldering fires in dwellings*", Suppression, Detection and Signaling Research and Applications Symposium (SupDet), San Antonio, Texas, USA, March 2016

[2] ISO 9705. *Fire tests - Full-scale room test for surface products*. International Organization for Standardization, Geneva, Switzerland, 1996.

[3] Cleary, T.G., *Results from a Full-Scale Smoke Alarm Sensitivity Study*, Fire Technology, May 2014, Vol. 50, Issue 3, pp 775-790 http://dx.doi.org/10.1007/s10694-010-0152-2

[4] ISO 13571:2007, *Life-threatening components of fire -- Guidelines for the estimation of time available for escape using fire data*.

[5] Cleary T.G., *A Study on the Performance of Current Smoke Alarms to the New Fire and Nuisance Tests Prescribed in ANSI/UL 217-2015*, Natl. Inst. Stand. Technol., Technical Note 1947, (2016) https://doi.org/10.6028/NIST.TN.1947

[6] Bukowski, R. W., Peacock, R. D., Averill, J. D., Cleary, T. G., Bryner, N. P., Walton W. D., Reneke, P. A., and Kuligowski, E. D., *Performance of Home Smoke Alarms, Analysis of the Response of Several Available Technologies in Residential Fire Settings*, Natl. Inst. Stand. Technol., Tech. Note 1455-1 (2008).

[7] ANSI/UL 217-2015: *Standard for Safety Smoke Alarms*, Underwriters Laboratories Inc., Northbrook, IL, 2015.

Polarized Light Scattering of Smoke Sources and Cooking Aerosols

Thomas Cleary, Amy Mensch
National Institute of Standards and Technology, Gaithersburg, MD, USA

Abstract

Light scattering data was gathered during experiments conducted in an ANSI/UL 217 test room constructed at the National Institute of Standards and Technology (NIST) to assess the performance of currently available smoke alarms. Smoldering and flaming fires along with cooking experiments were conducted. The light scattering device was configured to measure polarized light scattering characteristics of the fire smokes and cooking aerosols. Results are presented for forward scattering, polarization and asymmetry ratios. The results show a high degree of discrimination by a 90° polarization ratio between flaming soot and other smoldering smokes and cooking aerosols, and to a lesser degree discrimination by forward scattering and asymmetry ratios at the chosen angles.

Keywords: Smokes, cooking aerosols, light scattering

Introduction

The purpose of multiple measurement angles and/or light sources in smoke detection is to provide some discrimination of aerosols to distinguish smokes from non-fire sources. Weinert examined polarized light scattering from a number of fire and nuisance sources and showed a level of source discrimination using various measures [1]. Detectors and alarms that use multiple light scattering measures including different wavelengths, scattering angles and polarization states, perhaps combined with other sensor signals, may have the ability to distinguish between fire and non-fire conditions to a high degree. Given that new requirements in ANSI/UL 217-2015 [2] specifically require a cooking nuisance source test and apparently no current smoke alarms would pass the new requirements [3], there is an industry focus on detector modifications to meet the new standard. Data on the light scattering characteristics of the new fire and nuisance source tests and additional fire and nuisance aerosol sources may provide a foundation

Cleary, Thomas; Mensch, Amy.
"Polarized Light Scattering of Smoke Sources and Cooking Aerosols."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

for developing new discriminating detection schemes. Thus, NIST has begun to collect and analyze such data.

Experimental

Measurements were made with the NIST nephelometer/polarimeter [4] to gather polarized light scattering characteristics of fire smokes and nuisance source aerosols. The nephelometer section was configured to record vertically polarized light scattering intensities at two diode laser wavelengths, 638 nm and 980 nm, and five angles (15°, 22.5°, 45°, 90° and 135°) for each wavelength. In addition, horizontally polarized light scattering intensity at 90° for each wavelength was recorded. The acceptance angle for the scattered light reaching the detectors was about ± 3°. The data was acquired at 1 Hz to provide temporal resolution for the changing environment during each experiment. Neutral density filters were used to attenuate scattering signals to the measurement range of the photodetectors when needed. Additionally, laser light intensity (0°) was recorded and used to normalize the scattering intensities by the incident laser intensity. Figure 1 is a schematic of a cross-section for one beam. The aerosol flows through the central opening while the laser beam bisects the opening. What is not shown are polarization elements including a Glan-Thompson polarizer in front of the source beam to provide the incident polarization state, and ½ waveplates before the photodetectors to pass only scattered light with the desired polarization state.

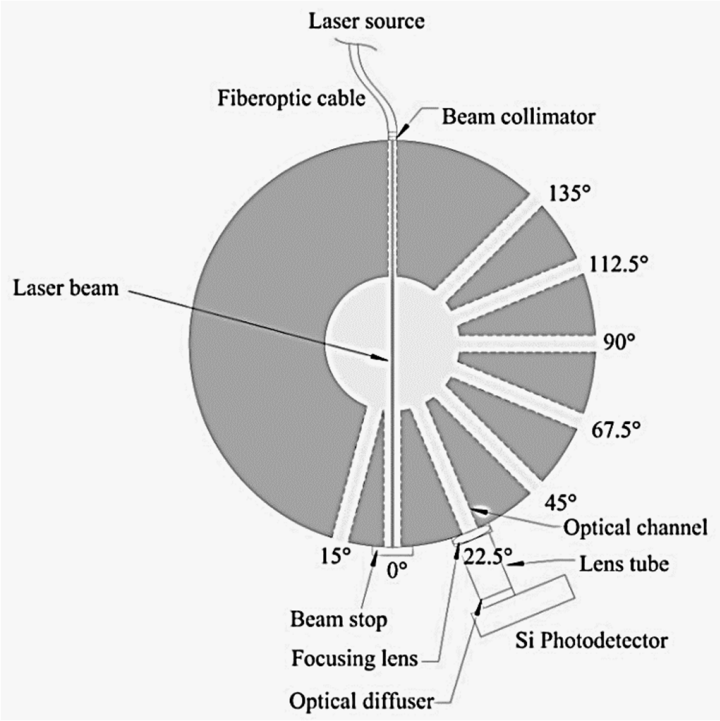


Figure 1. Schematic of a section of the nephelometer.

Cleary, Thomas; Mensch, Amy.
"Polarized Light Scattering of Smoke Sources and Cooking Aerosols."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

Following Weinert [1], forward scattering ratios (FR, I_{vv15^o} /$I_{vv22.5^o}$), asymmetry ratios (AR, I_{vv45^o} /I_{vv135^o}) and polarization ratios (PR, I_{hh90^o} /I_{vv90o}) were computed. Here, I is the scattering signal intensity, and the subscripts denote the horizontal (h) and vertical (v) polarization states of incident and scattered light, and the scattering angle. Initial calibrations were performed with nearly monodisperse di-ethyl-hexyl-sebacate (DEHS) particles of several aerodynamic diameters, from 0.18 µm to 1.0 µm, produced by a condensation/evaporation aerosol generator. The particle size distribution was measured with an electrical low pressure impactor and fitted to a log-normal distribution (d_g - geometric mean diameter, and σ_g - geometric standard deviation). The relative combined standard uncertainty in the mean diameter is estimated to be less than 10 %. Mie scattering calculations were performed with the results integrated over the size distribution and the acceptance angle of the nephelometer.

Results

The results are compared to Mie scattering calculations in Table 1. Some values were not tabulated which indicates either a low signal or a saturated signal of one of the photodetectors. The relative combined standard uncertainty for the computed ratios is estimated to be less than 10 % for the tabulated values.

Table 1. Measured size distributions and corresponding measured and computed scattering ratios for DEHS particles.

d_g (µm)	σ_g	$FR_{638\ nm}$ Measured/ Computed	$PR_{638\ nm}$ Measured/Computed	$AR_{638\ nm}$ Measured/Computed
0.18	1.49	- /1.15	0.013/0.19	3.64/8.60
0.26	1.39	1.03/1.17	0.065/0.34	6.98/14.5
0.30	1.26	1.06/1.11	0.110/0.29	11.2/16.2
0.39	1.28	1.09/1.22	0.660/0.78	29.2/26.3
0.52	1.27	1.16/1.39	- /1.01	14.4/18.6
0.66	1.25	1.30/1.61	- /1.10	24.0/8.54
0.83	1.25	1.68/1.98	- /1.33	5.19/5.29
1.02	1.30	2.35/2.40	- /1.40	9.84/6.31
d_g (µm)	σ_g	$FR_{980\ nm}$ Measured/ Computed	$PR_{980\ nm}$ Measured/Computed	$AR_{980\ nm}$ Measured/Computed
0.18	1.49	- /1.07	0.025/0.051	1.16/3.70
0.26	1.39	- /1.08	0.016/0.073	1.97/4.92
0.30	1.26	- /1.05	0.011/0.025	2.21/3.60
0.39	1.28	- /1.09	0.011/0.15	3.92/9.39
0.52	1.27	- /1.15	0.030/0.50	14.1/22.9
0.66	1.25	1.26/1.24	0.125/0.94	- /27.5
0.83	1.25	1.54/1.39	0.182/1.00	12.9/18.5
1.02	1.30	1.64/1.69	- /1.16	8.60/7.83

Cleary, Thomas; Mensch, Amy.
"Polarized Light Scattering of Smoke Sources and Cooking Aerosols."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

Figures 2-4 are plots comparing the various measured and computed ratios for the 638 nm wavelength beam. The measured and computed values follow the same trends with the exception of polarization ratios for the 980 nm wavelength beam. The trend is the same, but with a difference of a factor of about 10 to 20. This could indicate alignment issues.

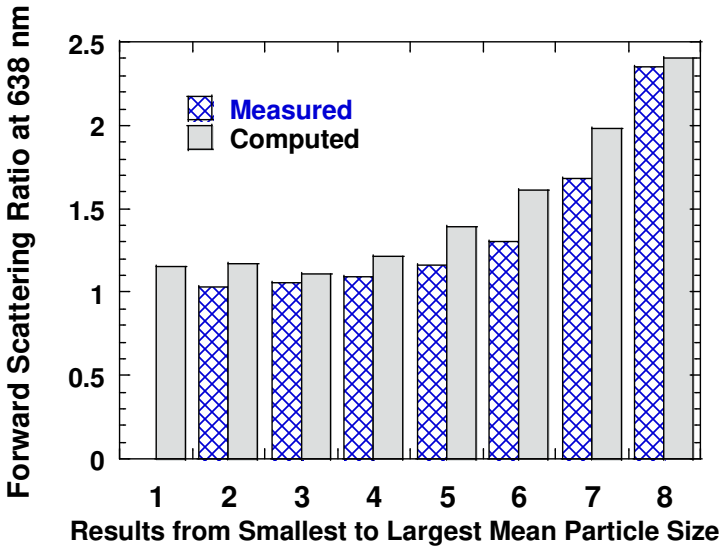


Figure 2. Results of the forward scattering ratios of DEHS particles.

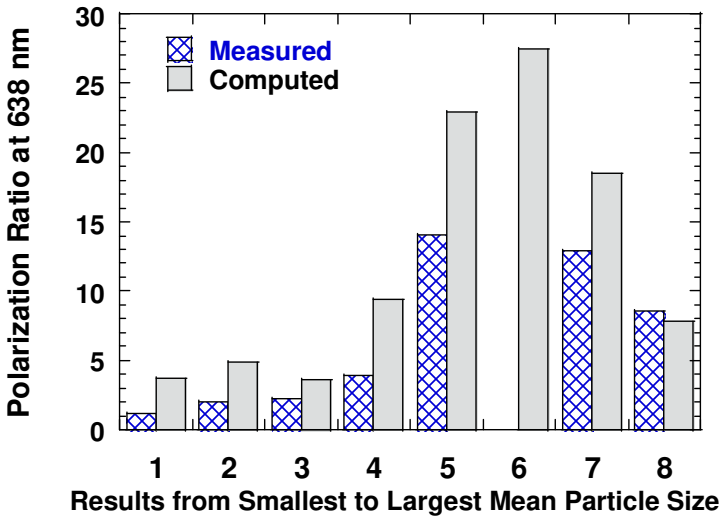


Figure 3. Results of the polarization ratios of DEHS particles.

Cleary, Thomas; Mensch, Amy.
"Polarized Light Scattering of Smoke Sources and Cooking Aerosols."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

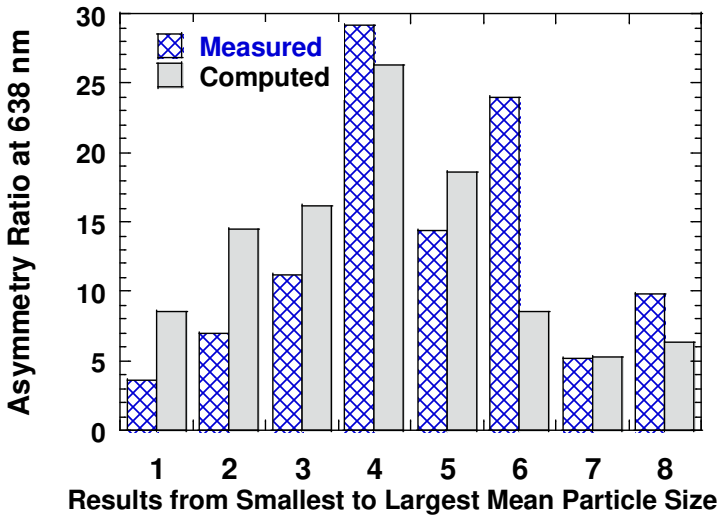


Figure 4. Results of the asymmetry ratios of DEHS particles.

Given the uncertainty in alignment and particle size, measured and computed values compare favorably, thus the instrument configuration provides realistic estimates of light scattering ratios except $PR_{980\ nm}$.

Aerosol samples from full-scale ANSI/UL 217-2015 room experiments were directed to the nephelometer/polarimeter. The flaming sources included polyurethane foam, a heptane/toluene pool and shredded copy paper. The smoldering sources included polyurethane foam and wood blocks on a hot plate. The cooking nuisance sources included broiling hamburgers, frying hamburger, stir-frying vegetables and heating cooking oil. The flaming and smoldering polyurethane foam and broiling hamburger experiments were conducted in the manner following ANSI/UL 217-2015 as described in reference [3].

Figures 5 and 6 show comparisons of smokes and cooking aerosols between ceiling beam obscuration and 45° forward light scattering at a wavelength of 638 nm normalized by the incident beam intensity. The difference between scattering and obscuration for the flaming foam smoke and the smoldering and cooking smokes is indicative of the relatively large absorption coefficient of black soot compared to the other sources. Table 2 shows the calculated ratios for the two wavelengths of the various sources. Values were averaged over an obscuration range indicative of smoke alarm activation concentration. Figures 7 and 8 are plots of the values for polarization ratio and asymmetry ratio at 638 nm wavelength. The flaming foam and heptane/toluene pool fire sooty smokes are easily discriminated from the other sources. However, the smoldering smokes and flaming paper smoke are not clearly distinguished from the cooking aerosols.

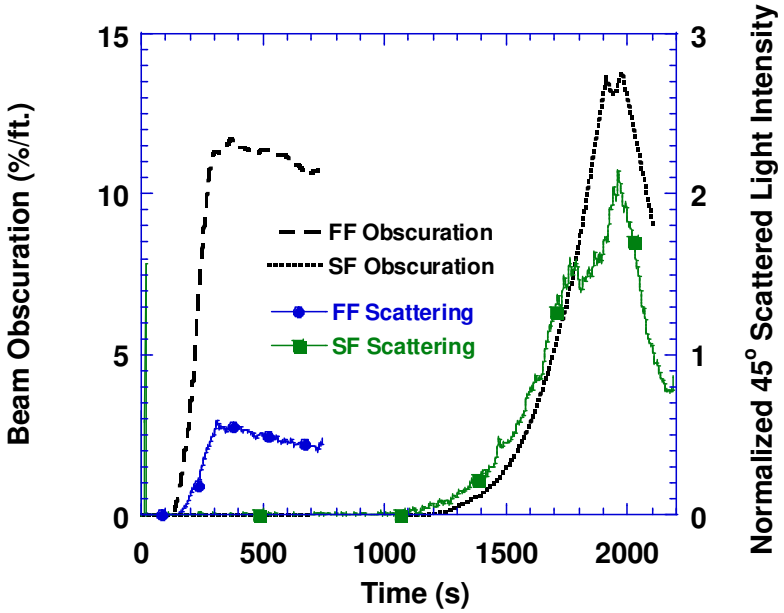


Figure 5. Ceiling beam obscuration and forward scattering signal for flaming polyurethane foam (FF) and smoldering foam (SF).

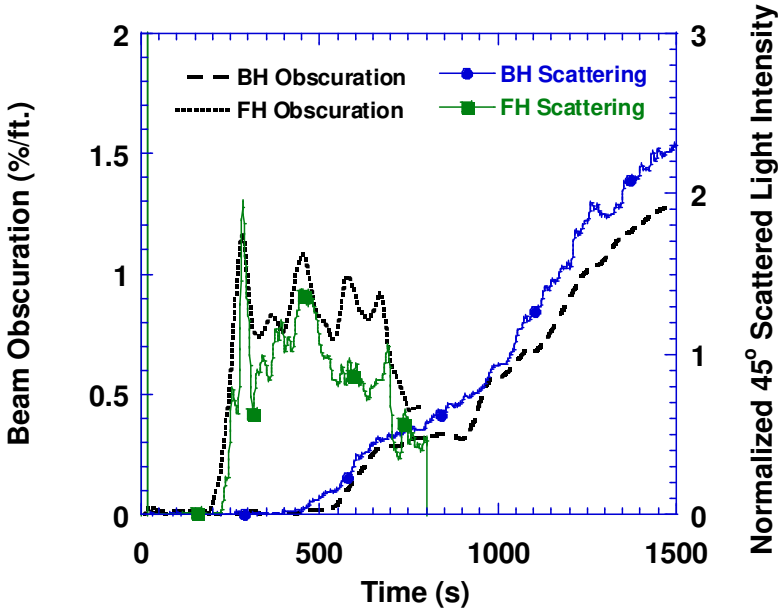


Figure 5. Ceiling beam obscuration and forward scattering signal for frying hamburger (FH) and broiling hamburgers (BH).

Cleary, Thomas; Mensch, Amy.
"Polarized Light Scattering of Smoke Sources and Cooking Aerosols."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

Table 2. Tabulated values of scattering ratios averaged over the indicated beam obscuration range for the various sources.

Source	Obsc. Range (%/ft.)	PR_{638} Ratio, SD*	AR_{638} Ratio, SD	PR_{980} Ratio, SD	AR_{980} Ratio, SD
Flaming Foam (FF)	2-6	0.014, 0.005	4.59, 0.19		3.80, 0.27
Heptane/ Toluene (H/T)	4-8	0.012, 0.001	4.65, 0.03		3.89, 0.01
Flaming Paper (FP)	1.5 - 2	0.23, 0.03	9.26, 1.13	0.44, 0.24	7.39, 0.29
Smoldering Foam (SF)	2-4	0.25, 0.01	15.0, 0.4	0.44, 0.04	4.35, 0.50
Smoldering Wood (SW)	0.5 – 0.75	0.27, 0.06	13.9, 2.5		
Broiling Hamburger (BH)	0.5 - 1.5	0.17, 0.01	9.98, 0.19	0.19, 0.01	5.45, 0.22
Frying Hamburger (FH)	1	0.37, 0.02	17.9, 0.4	0.28, 0.02	7.18, 0.63
Stir-frying Vegetables (SV)	1-1.5	0.39, 0.04	15.8, 0.7	0.36, 0.03	7.85, 0.72
Cooking Oil (Oil)	0.3 -0.7	0.27, 0.04	20.4, 1.0	0.23, 0.06	7.22, 0.50

* SD – standard deviation of the ratio over the obscuration range

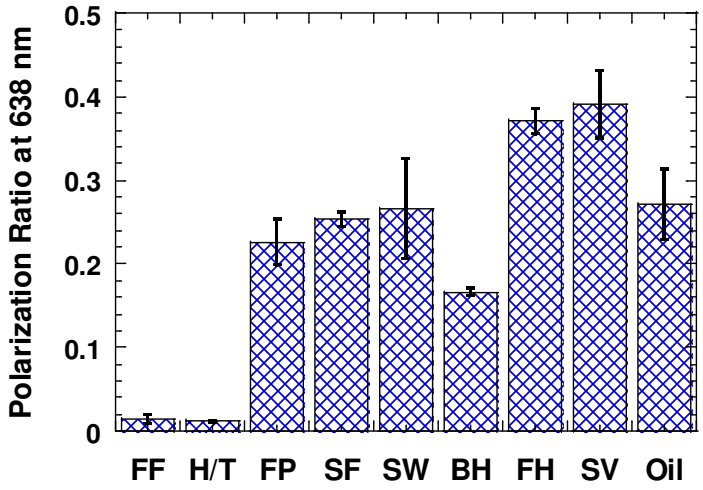


Figure 7. Polarization ratios for various sources, error bar indicates standard deviation.

Cleary, Thomas; Mensch, Amy.
"Polarized Light Scattering of Smoke Sources and Cooking Aerosols."
Paper presented at 16th International Conference on Automatic Fire Detection, AUBE '17 / Suppression, Detection and Signaling Research and Applications Conference, SUPDET 2017, College Park, MD, United States. September 12, 2017 - September 14, 2017.

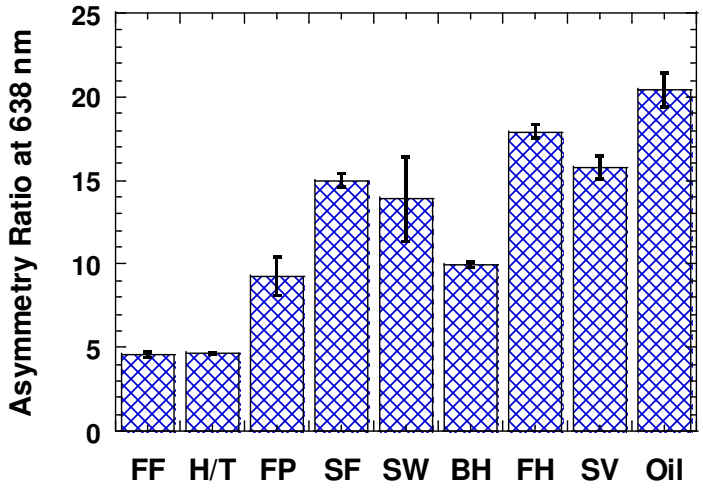


Figure 8. Asymmetry ratios for various sources, error bar indicates standard deviation.

Conclusions

The results show a high degree of discrimination between flaming soot and other smoldering and cooking aerosols considering a 90° polarization ratio, and a lesser degree of discrimination considering forward scattering and asymmetry ratios at the chosen angles similar to measurements conducted by Weinert.

Acknowledgements and Disclaimer

This research was funded in part by the U.S. Consumer Product Safety Commission. Official contribution of the U.S. Government; not subject to copyright in the United States.

References

[1] Weinert D., [2003] *Light Scattering by Smoke and Nuisance Aerosols*, Ph.D., Victoria University of Technology, Australia.

[2] *ANSI/UL 217-2015: Standard for Safety for Smoke Alarms*, Underwriters Laboratories Inc., Northbrook, IL, 2015.

[3] Cleary T.G., *A Study on the Performance of Current Smoke Alarms to the New Fire and Nuisance Tests Prescribed in ANSI/UL 217-2015*, Natl. Inst. Stand. Technol., Technical Note 1947, (2016) https://doi.org/10.6028/NIST.TN.1947

[4] Cleary, T.G., and M. Zarzecki, "*A Nephelometer/Polarimeter for Characterizing Smoke and Nuisance Alarm Sources*," Proceedings of the 15th International Conference on Automatic Fire Detection AUBE '14, Duisburg, Germany, October 14-16, 2014.

Modulated Photocurrent Measurements in Double Junction Solar Cells

Nicolás Márquez Peraca and Behrang H. Hamadani

National Institute of Standards and Technology, Gaithersburg, Maryland, 20899, United States

***Abstract* — Frequency dependent external quantum efficiency (EQE) measurements were performed on double junction solar cells by a custom-designed system consisting of an array of various monochromatic LEDs. LEDs were operated both at constant intensity and pulsed at various frequencies to explore the frequency response of each junction under various conditions. An equivalent circuit model, incorporating the effects of shunt resistances, junction capacitances, optical light coupling and the series resistance was then used to explain the various features and findings obtained from these measurements.**

I. Introduction

With significant improvements in design, fabrication, and performance of multijunction solar cells [1,2], it becomes necessary to establish more advanced opto-electronic characterization techniques to explore the characteristics of these solar cells. In recent years, extensive light bias and voltage bias dependent external quantum efficiency (EQE) measurements have been performed to elucidate artifacts and phenomena such as low shunt resistance effects [3,5-8,10,11], reverse breakdown voltage [3,4], light coupling between junctions [9,10,12], etc. in these devices. Most EQE measurements are performed using a differential spectral response system where a monochromatic light source incident upon the cell is chopped at a certain frequency creating an AC signal in the measurement junction of interest, while a DC light bias is applied to the other junctions. Although there has been much work discussing the effects observed under these circumstances, very little work has been dedicated to the frequency response of the AC photocurrent extracted from the current limited junction. One can think of this type of measurement as a frequency-dependent EQE, since the internal junction capacitances and resistances of each junction affect the extracted photocurrent magnitude and phase in the frequency domain.

In this work, we describe the result of our modulated photocurrent measurements in a simple double junction solar cell and show that an equivalent circuit model can be used to describe the unique features observed in both the amplitude and the phase response of the normalized photocurrent or the EQE of these solar cells. In particular, it is demonstrated that EQE shows a significant frequency dependence based on each junction's bias current and capacitive effects.

II. Experimental Details

A diagram of the experimental setup used for performing the modulated photocurrent measurements is shown in Fig. 1. A function generator is used in conjunction with a custom current amplifier to provide a pulsed AC signal to the LED array, while an LED controller provides the DC input. A solid glass light pipe in the form of a frustum is mounted in front of the LED array, which allows for a uniform illumination spot at the sample location (the exit port of the light pipe) for each LED type used. The cell's output is connected to a high-speed current to voltage pre-amplifier, which in turn is connected to a lock-in amplifier, providing amplitude and relative phase of the signal. This lock-in is synchronized with the function generator, and the whole system is controlled and automated by a computer program. Amplitude and phase dependence of the photocurrent on the frequency of the modulated light can then be found by changing the pulsed LED frequency. To provide stable operation of the LEDs, a water chiller is used to cool down the LED array plate to approximately 15 °C, as shown in Fig. 1.

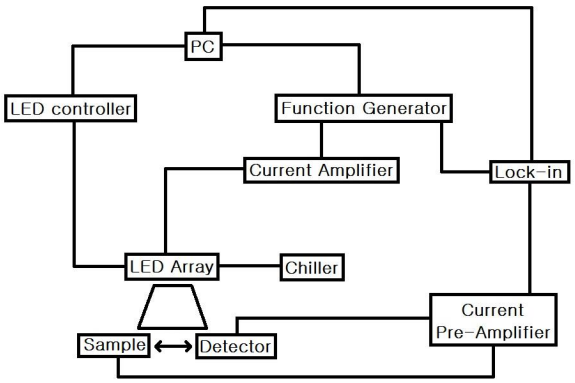


Fig. 1 Experimental setup used for the measurements. Both an AC and DC source are taken as an input of the LED array, which illuminates the sample through a quartz light pipe. A high-speed lock-in amplifier measures the amplitude and relative phase of the signal, and those values are then recorded on a computer.

Fig. 2. shows an actual photo of the optical segment of the setup. The LED array plate can be seen on the left side of the image. It has 12 LEDs of different wavelengths ranging from 460 nm to 928 nm. In this case, both the 460 nm and 623 nm LEDs are turned on, producing a purple color on the sample mounting plate due to the homogenizing of the blue and the red colors passing through the light pipe. The light pipe is in the center, encased in a 3D printed holder.

Marquez Peraca, Nicolas; Hamadani, Behrang.
"Modulated Photocurrent Measurements in Double Junction Solar Cells."
Paper presented at 44th IEEE Photovoltaic Specialists Conference (PVSC 2017), Washington, D.C., United States. June 25, 2017 - June 30, 2017.

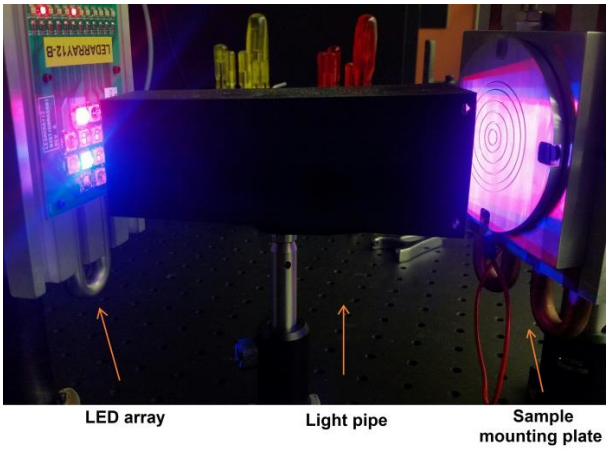


Fig. 2. Photo of the experimental setup. Both 460 nm (blue) and 623 nm (red) LEDs are turned on, producing a purple color at the sample mounting plate.

The solar cell used for this study was a *GaInP/GaAs* cell, with an illuminated active area of 0.2533 cm^2. The top *GaInP* junction is 0.9 μm thick with a bandgap around 1.84 eV, and the bottom junction is 3.5 μm thick. The active region of the top and bottom junctions were characterized by performing spectral response measurements on the cell by use of a monochromator-based system (see Fig. 3). Then, 460 nm and 850 nm LEDs were selected for the setup in Fig. 1, the former used as the pulsed light and the latter as bias light. Throughout the measurements the intensity of the pulsed 460 nm LED was kept fixed at 0.5 W/m^2.

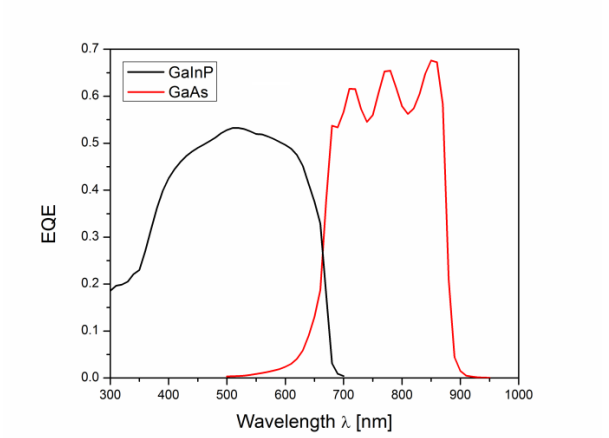


Fig. 3. External quantum efficiency as a function of wavelength for the *GaInP/GaAs* cell, obtained by performing spectral response measurements on the cell by use of a monochromator-based system.

III. Theoretical Model

An equivalent circuit for the AC light excitation measurements on the double junction solar cell can be seen in Fig. 4. $\tilde{I}_T$ and $\tilde{I}_B$ represent the AC currents generated in each junction, R_T and R_B their dynamic resistances (which might depend on the DC light bias current), and C_T and C_B the depletion region capacitances. The dependent current source $\eta_1 \tilde{I}_r$ models any possible light coupling from the top junction to the bottom. The general solution for the circuit-extracted current, $\tilde{I}_{SC}$ in this model can be found in [9]. In the case where the pulsed light is applied to the top junction while the bottom junction is DC light biased, i.e., similar to EQE measurement conditions for the top junction, this result simplifies to:

$$\frac{\tilde{I}_{SC}}{\tilde{I}_T} = \frac{Z_T}{Z_T + Z_B + R_S} = X(\omega) + jY(\omega), \qquad (1)$$

when there is no light coupling from the top junction to the bottom, and:

$$Z_i = \frac{R_i}{1 + j\omega C_i R_i}, \qquad (2)$$

$$R_B = \frac{nk_BT}{q}\frac{1}{I_B} = \frac{nV_T}{I_B}, \qquad (3)$$

where n is the diode ideality factor, k_BT/q is the thermal voltage V_T ($\approx$ 25 mV at room temperature), Z_i (for i=T, B) is the complex impedance for the top and bottom junction, I_B is the DC current generated in the bottom junction, and $X(\omega), Y(\omega)$ are the real and imaginary parts of $\tilde{I}_{SC}/\tilde{I}_T$, respectively. It is noted that the ratio $\tilde{I}_{SC}/\tilde{I}_T$ actually represents the internal quantum efficiency (IQE) of this cell because $\tilde{I}_T$, the AC photocurrent generated in the top junction, is proportional to the modulated light intensity, $\tilde{E}_T$. Therefore, the ratio $\tilde{I}_{SC}/\tilde{I}_T \propto \tilde{I}_{SC}/\tilde{E}_T = \tilde{R}_T \propto IQE$, $\tilde{R}_T$ being the internal spectral responsivity of the junction. We have multiplied this value by a fixed constant before comparing the model to the experimental data to include reflectance effects, so that it can represent the EQE at the excitation wavelength probed.

It can be easily seen from (2) that in the low frequency limit where $\omega \to 0$, $Z_i \to R_i$ and so:

$$\frac{\tilde{I}_{SC}}{\tilde{I}_T} \to \frac{R_T}{R_T + R_B + R_S} \approx 1 \qquad (4)$$

where the last step follows from the approximation $R_T >> R_B, R_S$ when the top cell is in reverse bias.

On the other hand, in the high frequency limit $\omega \to +\infty$, $Z_i \to 1/j\omega C_i$ and if we consider a negligible series resistance, then we find:

Marquez Peraca, Nicolas; Hamadani, Behrang.
"Modulated Photocurrent Measurements in Double Junction Solar Cells."
Paper presented at 44th IEEE Photovoltaic Specialists Conference (PVSC 2017), Washington, D.C., United States. June 25, 2017 - June 30, 2017.

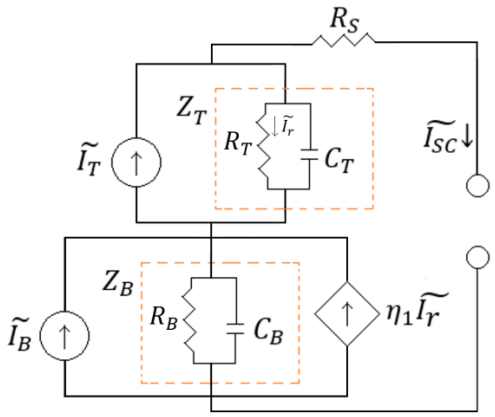


Fig. 4. Equivalent circuit for the AC measurements performed on the double junction cell. The two AC sources represent the currents generated on the junctions, whilst the dependent current source accounts for any possible light coupling.

$$\frac{\tilde{I}_{SC}}{\tilde{I}_T} = \frac{1}{1+\frac{Z_B}{Z_T}} \to \frac{1}{1+\frac{C_T}{C_B}} \quad . \qquad (5)$$

The ratio C_T / C_B can then be obtained from the EQE(ω) plot.

The phase $\theta(\omega) = \text{Arctan}(Y(\omega)/X(\omega))$ presents a resonant behavior and, as it can be seen by taking the quotient of the imaginary and real parts of (1), when there is no series resistance present it goes to zero both in the high and low frequency limits. The value for which this resonant peak happens is found to be:

$$\omega_{\min} = \frac{\sqrt{1+R_B/R_T}}{R_B C_B \sqrt{1+\frac{C_T}{C_B}}} \approx \frac{1}{R_B C_B} \propto I_B \ , \qquad (6)$$

a result that can be verified by setting the first derivative of $\theta(\omega)$ to zero, and where the last step follows from (3). Furthermore, the real part of $\tilde{I}_{SC}/\tilde{I}_T$ evaluated at this frequency is equal to:

$$\left.\frac{\tilde{I}_{SC}}{\tilde{I}_T}\right|_{\omega=\omega_{\min}} = 2\left[\left(\frac{\tilde{I}_{SC}}{\tilde{I}_T}(\omega \to 0)\right)^{-1} + \left(\frac{\tilde{I}_{SC}}{\tilde{I}_T}(\omega \to +\infty)\right)^{-1}\right]^{-1} \qquad (7)$$

On the other hand, a non-zero series resistance causes a drop in the amplitude from the value in (5) to zero, and makes the phase rotate from 0 to -90° in the high frequency limit. Even so, in the region where the condition $\omega_{\min}(I_B) < 1/R_S C_T$ is met the previous analysis continues to be approximately valid.

These features can be seen in Fig. 5, where the left Y-axis represents the magnitude $IQE(\omega) = \sqrt{X(\omega)^2 + Y(\omega)^2}$ and the right Y-axis the phase $\theta(\omega)$, expressed in degrees. Here R_S was kept fixed at 90 Ω (see next section) while I_B was increased from $0.1\,\mu A$ to $100\,mA$ for exemplification purposes. The curve for $I_B = 100\,mA$ shows the case where $1/R_S C_T < \omega_{\min}$, and so no resonance is present in the phase plot and the IQE shows a sudden drop to zero at $\omega \sim 1/R_S C_T$.

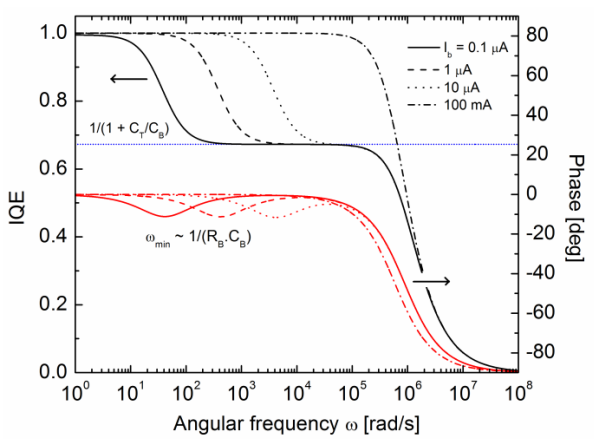


Fig. 5. Predicted internal quantum efficiency as a function of frequency. Here R_S was kept fixed at 90Ω, while I_B was increased from $0.1\mu A$ to $100mA$. At the frequency $\omega_{\min}$, both a resonant behavior for the phase and a sudden drop in the IQE to the value in (5) occur. The high and low frequency limits of (4) and (5) can also be seen in this figure.

IV. Discussion and Results

Starting from the amplitude and phase data obtained from the lock-in measurements in the experimental setup, both for the solar cell and the reference detector, the external quantum efficiency (EQE) and net phase can be calculated. Fig. 6 shows the results obtained from these measurements (scatter points), as well as the model predictions of (1) (solid lines). The left Y-axis represents the external quantum efficiency and the right Y-axis is the phase, expressed in degrees. Both the model predictions and the measurements were scaled to the EQE value obtained from the monochromator setup at 460 nm, whilst a fixed ≈ 1° was subtracted from the phase data to account for the unphysical non-zero phase at low frequencies related to a small phase lag in the instrumentation. The series resistance used in the model was 90 Ω : 50 Ω corresponding to the pre-amplifier's input impedance (from the specification data) and 40 Ω obtained through I-V measurements from the cell itself. Setting 1 to Setting 4 correspond to different DC light bias conditions which, expressed in terms of the LED controller current values, are 1, 2, 5, and 10 mA, respectively.

Marquez Peraca, Nicolas; Hamadani, Behrang.
"Modulated Photocurrent Measurements in Double Junction Solar Cells."
Paper presented at 44th IEEE Photovoltaic Specialists Conference (PVSC 2017), Washington, D.C., United States. June 25, 2017 - June 30, 2017.

As Fig. 6 shows, the EQE drop occurs around the same frequency where the phase minimum happens and, as expected from (6), it shifts towards higher frequencies when the DC current generated on the bottom junction increases. This effect suggests that when performing monochromator-based differential spectral response measurements for determining the steady-state EQE of the cell, the chopper's frequency should satisfy the condition $\omega_{meas} < I_B / nV_T C_B$. Otherwise, one risks underestimating the correct magnitude of the EQE for a given junction. In general, it is recommended to perform EQE measurements under the lowest frequencies possible, particularly when light bias conditions are low.

The little discrepancy between the model and the measurements in the high frequency region in Fig. 6 is explained by the fact that our current to voltage pre-amplifier has a strong bandwidth dependence with the source capacitance. For the measured cells having capacitances around 30 nF, the bandwidth drops to approximately 100 kHz to 200 kHz from its maximum value of ≈ 100 MHz.

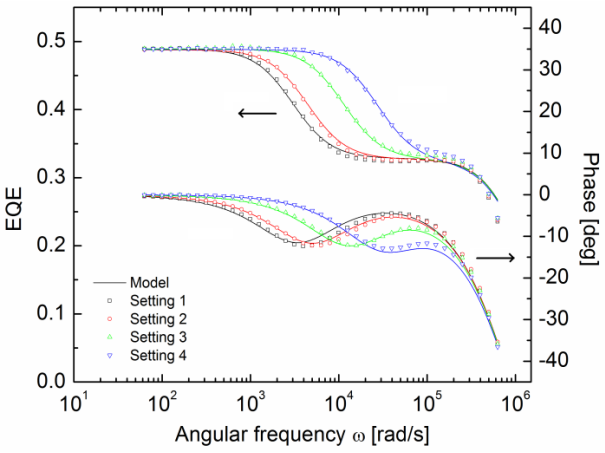


Fig. 6. Results obtained from the measurements superposed to the theoretical model predictions, both being scaled to the EQE reported by the monochromator system. As shown, the optimal measurement frequency depends on the light bias intensity and on the bottom junction capacitance (see text). The values used for the fits were initially estimated from published works, and then adjusted through the model to obtain $C_T = 18.7$ nF , $C_B = 38.4$ nF , $R_S = 90\ \Omega$, $R_T \approx 10^8\ \Omega$, and $I_B = 8.1\ \mu$A, 11.93 μA, 30.51 μA, 74.78 μA for Settings 1-4, respectively.

V. Conclusions

Measurements of the amplitude and phase dependence of the external quantum efficiency on frequency for a double junction cell were performed and compared against the predictions of an equivalent circuit model. It was shown that in general the optimal measurement frequency will depend both on the light bias intensity levels and the capacitance of the junction that is in forward bias. Our recommendation is to use a measurement frequency as low as possible, while increasing the light bias. The frequency-dependent photocurrent measurements also allow for the determination of the internal capacitances and resistances of each junction by fitting the described model to a large set of data.

Acknowledgement

The authors would like to thank Dr. Daniel Friedman of NREL for graciously providing the solar cells used in this study. N. M. P. would like to thank the Solar Energy Laboratory (LES, Uruguay) and the Technological Laboratory of Uruguay (LATU) for their support of the research projects on which he was selected to participate, and to NIST for their hospitality throughout this stay. The authors also gratefully acknowledge the support of the NIST International and Academic Affairs Office.

References

[1] H. Yoon *et al.*, "Recent advances in high-efficiency III-V multi-junction solar cells for space applications: Ultra triple junction qualification," in *Progress in Photovoltaics: Research and Applications*, 2005, vol. 13, no. 2, pp. 133–139.

[2] M. A. Green, K. Emery, Y. Hishikawa, W. Warta, and E. D. Dunlop, "Solar cell efficiency tables (version 48)," *Prog. Photovoltaics Res. Appl.*, vol. 24, no. 7, pp. 905–913, 2016.

[3] J. P. Babaro, K. G. West, and B. H. Hamadani, "Spectral response measurements of multijunction solar cells with low shunt resistance and breakdown voltages," *Energy Sci. Eng.*, vol. 4, no. 6, pp. 372–382, 2016.

[4] M. Meusel, C. Baur, G. Létay, A. W. Bett, W. Warta, and E. Fernandez, "Spectral Response Measurements of Monolithic GaInP/Ga(In)As/Ge Triple-Junction Solar Cells: Measurement Artifacts and their Explanation," *Prog. Photovoltaics Res. Appl.*, vol. 11, no. 8, pp. 499–514, 2003.

[5] J.-J. Li, S. H. Lim, and Y.-H. Zhang, "A novel method to eliminate the measurement artifacts of external quantum efficiency of multi-junction solar cells caused by the shunt effect," *Proc.SPIE*, vol. 8256, pp. 825616–825623, 2012.

[6] J.-J. Li and Y.-H. Zhang, "Elimination of Artifacts in External Quantum Efficiency Measurements for Multijunction Solar Cells Using a Pulsed Light Bias," *IEEE J. Photovoltaics*, vol. 3, no. 1, pp. 364–369, 2013.

[7] V. Paraskeva, M. Hadjipanayi, M. Norton, M. Pravettoni, and G. E. Georghiou, "Voltage and light bias dependent quantum efficiency measurements of GaInP/GaInAs/Ge triple junction devices," *Sol. Energy Mater. Sol. Cells*, vol. 116, pp. 55–60, 2013.

[8] G. Siefer, C. Baur, and A. W. Bett, "External quantum efficiency measurements of Germanium bottom subcells: Measurement artifacts and correction procedures," in *Conference Record of the IEEE Photovoltaic Specialists Conference*, 2010, pp. 704–707.

[9] M. A. Steiner, S. R. Kurtz, J. F. Geisz, W. E. McMahon, and J. M. Olson, "Using phase effects to understand measurements of the quantum efficiency and related

luminescent coupling in a multijunction solar cell," *IEEE J. Photovoltaics*, vol. 2, no. 4, pp. 424–433, 2012.
[10] J.-J. Li, S. H. Lim, C. R. Allen, D. Ding, and Y.-H. Zhang, "Combined Effects of Shunt and Luminescence Coupling on External Quantum Efficiency Measurements of Multijunction Solar Cells," *IEEE J. Photovoltaics*, vol. 1, no. 2, pp. 225–230, 2011.
[11] M. Pravettoni, R. Galleano, A. Virtuani, H. Müllejans, and E. D. Dunlop, "Spectral response measurement of double-junction thin-film photovoltaic devices: the impact of shunt resistance and bias voltage," *Meas. Sci. Technol.*, vol. 22, no. 4, p. 45902, 2011.
[12] M. A. Steiner *et al.*, "Measuring IV curves and subcell photocurrents in the presence of luminescent coupling," *IEEE J. Photovoltaics*, vol. 3, no. 2, pp. 879–887, 2013.

Marquez Peraca, Nicolas; Hamadani, Behrang.
"Modulated Photocurrent Measurements in Double Junction Solar Cells."
Paper presented at 44th IEEE Photovoltaic Specialists Conference (PVSC 2017), Washington, D.C., United States. June 25, 2017 - June 30, 2017.

Solid Freeform Fabrication 2017: Proceedings of the 28th Annual International
Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference

CONTINUOUS LASER SCAN STRATEGY FOR FASTER BUILD SPEEDS IN LASER POWDER BED FUSION SYSTEM

H. Yeung[1], B. Lane[1], J. Fox[1], F. Kim[1], J. Heigel[1], J. Neira[2]

[1]Engineering Laboratory, [2]Physical Measurement Laboratory,

National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

Research has shown significant influence of laser scan strategy on various part qualities in the laser powder bed fusion additive manufacturing process. The National Institute of Standards and Technology developed the Additive Manufacturing Metrology Testbed, which provides open architecture for flexible control and monitoring during a laser powder bed fusion additive manufacturing process. This allows extended control of scan strategies, including control of laser power and speed within each scan line. A 'continuous' scan strategy can reduce build times and improve throughput by negating the need to turn the laser off between scan tracks (e.g., sky-writing). Also, less frequent laser power interruption can potentially improve the melt-pool continuity. Multiple experiments are performed utilizing the continuous and traditional scan strategies, and comparisons are made between build time and measured melt-pool qualities.

Introduction

Laser powder bed fusion (LPBF) is an additive manufacturing (AM) process in which a focused, high power laser selectively melts geometric patterns into layers of metal powder, ultimately building a near fully dense freeform part. The LPBF fabricated part quality is determined by many process parameters [1], such as the laser scan strategies (position, power, and velocity) and their respective synchronization, in conjunction with powder layer parameters (material, relative density, layer height, etc.). Varying the relative combination of these parameters can introduce known defects that plague LPBF parts. Pores, for example, have been attributed to various phenomena related to the power-velocity attributes or scan strategies (e.g., keyholing and collapse at high laser energy densities [2], or insufficient re-melting of adjacent scan vectors due to wide hatch spacing [3,4]). For example, Khairallah et al. noted that turning the laser off at the end of a scan vector can potentially cause pores to be trapped under the rapidly solidified melt pool, and recommended laser power decreased at these locations [5]. More adequately controlled velocity or power profiles along each scan vector can reduce probability of pore formation, or provide a parametric space for other property optimization. Apart from solidification physics at the end of a single vector, the general size, shape, and timing of a laser scanning raster pattern are known to affect the melt pool thermal history of the part, thus the resulting local and global residual stress and microstructure [6–8].

Though LPBF technologies are rapidly improving and maturing, there is still wide potential for research in melt pool, scan track, and layer formation process physics. To fully define scan strategies, a system needs to control both how to fill up the build areas with scan vectors (hatching), and position, power, and velocity of individual scan vectors. The National Institute of Standards and Technology (NIST) designed and built an open architecture Additive Manufacturing Metrology Testbed (AMMT). The control software developed for AMMT utilizes the full access of system parameters to enable quick deployment of different scan strategies. This work compares three scan strategies implemented on AMMT: continuous linear, continuous concentric, and constant build speed (sky-writing).

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Yeung, Ho; Lane, Brandon; Fox, Jason; Kim, Felix; Heigel, Jarred; Neira, Jorge.
"Continuous Laser Scan Strategy for Faster Build Speeds in Laser Powder Bed Fusion System."
Paper presented at The 28th Annual International Solid Freeform Fabrication Symposium, Austin, TX, United States. August 7, 2017 - August 10, 2017.

AMMT Scan Strategies

The AMMT is a facility at NIST which enables flexible and well-characterized monitoring and control of the laser powder bed fusion (LPBF) process [9,10]. AMMT control adapts an open platform design, using stereolithography (STL) files, G-code, and xy2-100 industrial standards for parts definition, path description, and galvanometer control, respectively. Figure 1 shows a generic software architecture for the AM system. The highlighted parts are modules implemented for AMMT. The step 3 'G-code Generator' converts a two-dimensional shape into scan paths described by G-code [11]. A standard G-code line, traditionally used for numerically controlled machines, describes a 'move' with destination point and speed. The LPBF-specific G-code version for NIST AMMT also includes laser power information. Step 4 'G-code Interpreter' translates this G-code into xy2-100 commands per the predefined interpretation mode. Xy2-100 commands [9,10] are digital positions updated every 10 μs to control a laser unit and galvanometer scanner. Therefore, the laser scan strategies are implemented by both 'G-code Generator' and 'G-code Interpreter'.

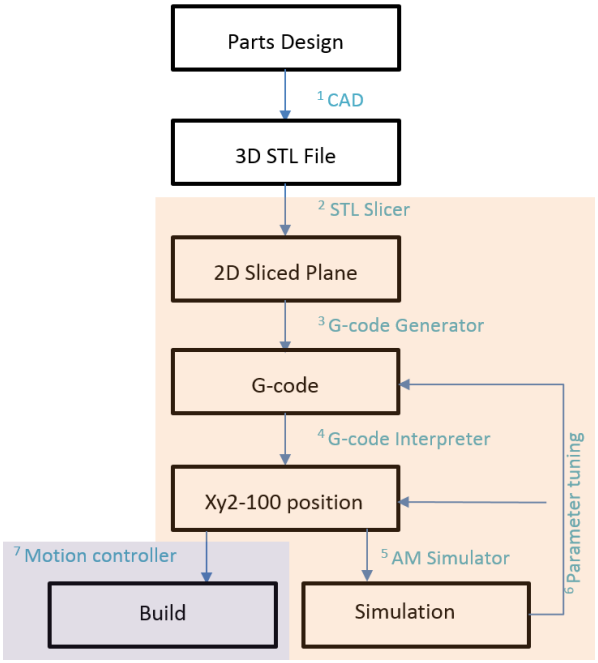


Figure 1: AMMT control design

G-code Generator

To build a layer, a two-dimensional shape must be filled up with scan vectors (hatching) which are spaced densely enough to melt the powder into a joint solidified region. The hatching pattern can be linear or concentric as shown in Figure 2a. The hatching lines can be scanned in an interleaved manner, or sequentially with alternative power as shown Figure 2b. A larger scan area can also be divided into smaller sub-regions (islands) and scanned in a certain sequence (Figure 2c). The combinations of above form the hatching strategy. G-code Generator fills a two-dimensional shape with scan vectors per the predefined hatching strategy, and generates G-code representing the scan vectors.

Yeung, Ho; Lane, Brandon; Fox, Jason; Kim, Felix; Heigel, Jarred; Neira, Jorge.
"Continuous Laser Scan Strategy for Faster Build Speeds in Laser Powder Bed Fusion System."
Paper presented at The 28th Annual International Solid Freeform Fabrication Symposium, Austin, TX, United States. August 7, 2017 - August 10, 2017.

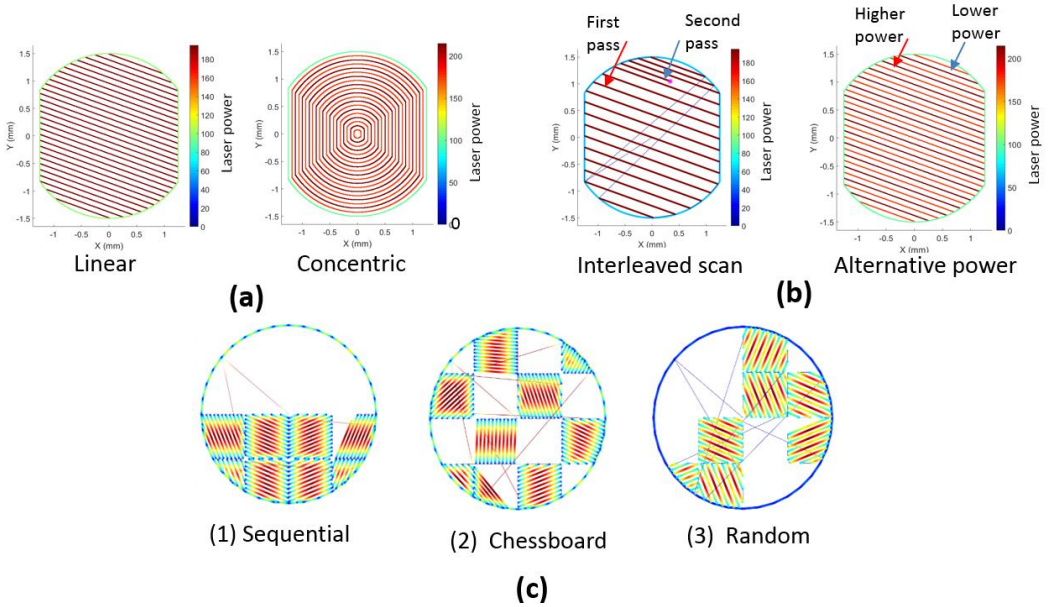


Figure 2: NIST AMMT hatching strategies: (a) Linear and concentric hatching patterns. (b) Interleaved scan and alternative power scan. (c) Islands with different scan sequence.

G-code Interpreter

The motion and power control of the laser spot within a move, as well as the transition between two sequential moves, is implemented in G-code Interpreter. Earlier work by the authors [9] proposed three laser path modes and three laser power modes. A summary is given below.

a) Laser path mode
   1. Exact stop mode – motion stops exactly at the end of each move with maximum allowable deceleration. If there is a subsequent move, motion will start immediately again with maximum allowable acceleration until it reaches the programmed speed, or until it needs to decelerate again.
   2. Constant build speed mode – motion is kept at constant speed for the whole move, if the laser power is on for that move (i.e., a build move). In order to achieve that, extra moves may be added before or after the original move to speed up or slow down the motion, but laser power is kept off for the added moves.
   3. Continuous mode - two sequential moves of different velocities are connected by an arc to allow a smooth transition of velocity. Deviation from the designed path within the maximum tolerance is allowed.

b) Laser power mode
   1. Constant power mode – power is kept constant at the programmed level during each build move, regardless of scan speed. If a subsequent move has a different power level, control sets the power to the new level.
   2. Constant density mode - power/speed ratio (power density) is kept at a predefined constant during each move. This constant is not necessarily the same for all moves.
   3. Thermal adjusted mode - power level is adjusted according to the predefined thermal properties of the building process.

Yeung, Ho; Lane, Brandon; Fox, Jason; Kim, Felix; Heigel, Jarred; Neira, Jorge.
"Continuous Laser Scan Strategy for Faster Build Speeds in Laser Powder Bed Fusion System."
Paper presented at The 28th Annual International Solid Freeform Fabrication Symposium, Austin, TX, United States. August 7, 2017 - August 10, 2017.

Experiment design

Different scan strategies can be programmed on AMMT by the combination of hatching strategies, laser path modes, and laser power modes. Seven experiments were designed. Figure 3 shows the actual laser scan tracks from each experiment. All experiments were conducted on a ground stainless-steel plate (100 mm x 100 mm x 6.4 mm) without powder. The scan speed is 500 mm/s, power is 200-watt constant, and hatch spacing is 200 µm. Wider spacing is used so individual tracks can be observed. Table 1 lists scan area (length x width) and scan strategy for each experiment. Experiment 1, 4, and 5 used constant build speed mode, also known as 'sky-writing' on some commercial LPBF systems, and are therefore marked as conventional and used as a base line for comparison. Others are marked as 'AMMT' in Table 1 as authors are not aware of any similar scan strategies on commercial systems, although this does not mean that similar concepts have not been explored before [12,13].

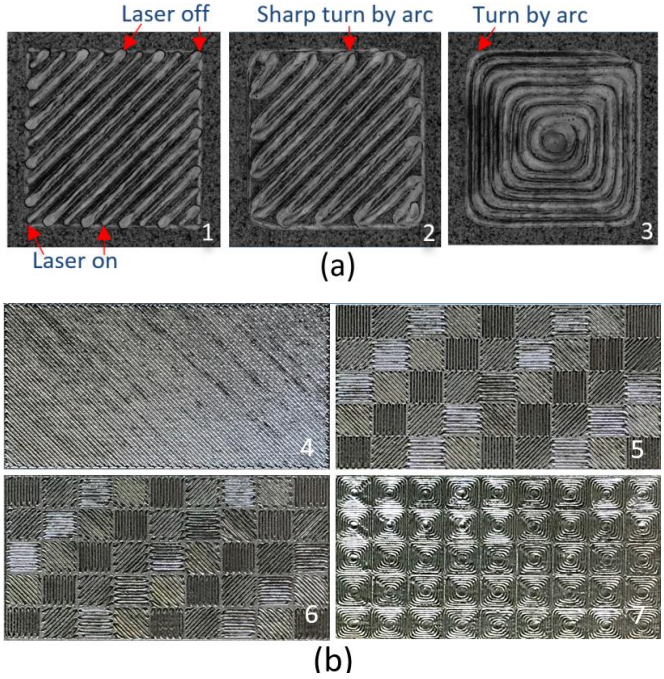


Figure 3. Scans on stainless steel plate with 500 mm/s speed and 200-watt constant laser power. (a) Single square scan experiments 1-3. (b) Island scan experiments 4-7.

Table 1. Experiment parameters

Exp #	Length x Width (mm)	Scan Strategies	Remarks
1	2.5 x 2.5	*Constant build speed*	Conventional
2	2.5 x 2.5	*Continuous linear*	AMMT
3	2.5 x 2.5	*Continuous concentric*	AMMT
4	20 x 10	*Constant build speed*	Conventional
5	20 x 10	*Constant build speed - island*	Conventional
6	20 x 10	*Continuous linear - island*	AMMT
7	20 x 10	*Continuous concentric - island*	AMMT

Yeung, Ho; Lane, Brandon; Fox, Jason; Kim, Felix; Heigel, Jarred; Neira, Jorge.
"Continuous Laser Scan Strategy for Faster Build Speeds in Laser Powder Bed Fusion System."
Paper presented at The 28th Annual International Solid Freeform Fabrication Symposium, Austin, TX, United States. August 7, 2017 - August 10, 2017.

Melt-pool Analysis

A key signature characteristic in LPBF AM processes is the melt-pool geometry. One can observe the melt-pool to study the influence of laser control on the process. In this study, in-situ high-speed coaxial imaging is used to measure the melt-pool area, and post-process confocal laser scanning microscopy is used to measure the melt-pool height.

Melt-pool width measurement

To measure melt-pool area, a high-speed camera was setup coaxially with the laser beam using a dichroic mirror, imaging lens, and filter. Emitted light from the melt-pool, which is filtered at 850 nm (40 nm bandwidth), is imaged on the camera sensor with nominal 1:1 magnification and 20 μm pixel size. The camera is set to 50 000 frames/s, 10 μs exposure time, 128 pixel x 128 pixel window, and 8-bit dynamic range (grayscale). The gray levels are used to relate to melt-pool dimensions [14]. Contours, representing isotherm lines, can be drawn on the raw melt-pool image to represent equal intensity (Figure 4). A contour with intensity of 50 digital levels was found to equate to the physical melt pool width based on the ex-situ measured scan track width via microscope inspection. This digital level contour is then used to infer the melt pool boundary from the high-speed images and to calculate melt pool dimensions and area.

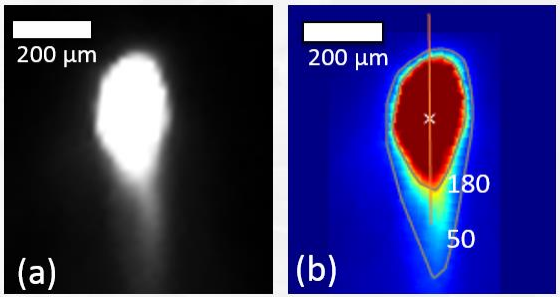


Figure 4. Melt-pool image analysis. (a) raw grayscale image. (b) processed image. Gray contour lines show different intensity levels (DL); red line shows melt-pool orientation.

Melt-pool height measurement

Confocal laser scanning microscopy enables the reconstruction of three-dimensional surfaces from a set of images obtained at different focal depths. A topography of the scan track such as Figure 5a can hence be created, with depth (height) represented by color. It is a very useful tool to visualize the effects of scan strategy on surface profile. From the topography, scan tracks are always bounded by valleys (Figure 5b), likely caused by surface tension of the molten metal. Once the two boundary valleys are identified, the melt-pool height can be defined as the peak between these two valleys (Figure 5c). Ideally the melt-pool width can also be determined by distance between the two valleys, but surface roughness or unsteady melt-pool behavior can sometimes create a local minimum (valley), too. Nevertheless, the local maximum is generally unique, making it a reliable approach for melt-pool height detection.

Yeung, Ho; Lane, Brandon; Fox, Jason; Kim, Felix; Heigel, Jarred; Neira, Jorge.
"Continuous Laser Scan Strategy for Faster Build Speeds in Laser Powder Bed Fusion System."
Paper presented at The 28th Annual International Solid Freeform Fabrication Symposium, Austin, TX, United States. August 7, 2017 - August 10, 2017.

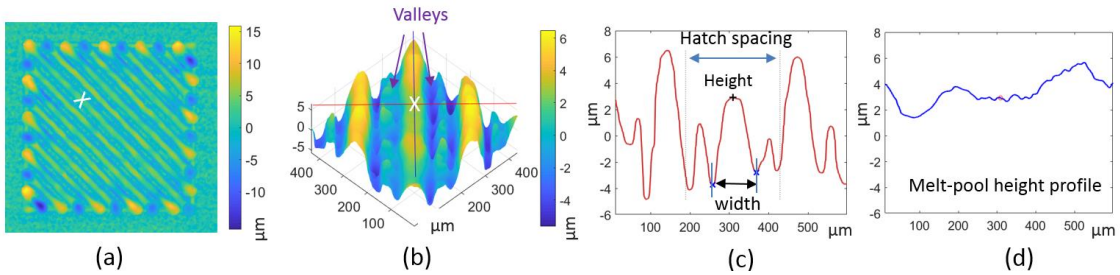


Figure 5 Melt-pool width and height detection from confocal microscopy. (a) Topography of the melt track from Exp. 1. (b) 400 μm x 400 μm three-dimensional plot centered at position marked by white x. (3) Cross section profile along the red line on three-dimensional plot. (d) Profile along the blue line.

Experiment results and discussion

Experiments 1-7 were conducted using AMMT on a bare metal (stainless steel) plate; no powder was added. The microscopic images of the scanned regions were shown in Figure 3. The processes were in-situ imaged by the high-speed coaxial camera at 50 000 frame/s with 10 μs exposure time. The images were processed to obtain melt-pool area, with the results plotted in Figure 6. The scanned regions were then examined by confocal microscope. The surface profiles are plotted in Figure 7 and Figure 8. The surface profile of an un-scanned region on the metal plate is also shown in Figure 7 as a reference.

Figure 6 shows the melt-pool area from the high-speed coaxial image analysis. Dark lines plot melt-pool areas for each individual frame, and red lines are the means for each experiment with one standard deviation. Table 2 lists mean, standard deviation, and build time specified for each experiment.

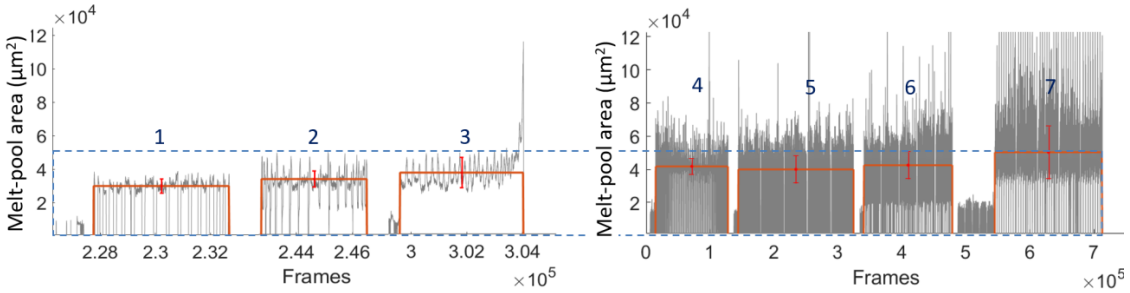


Figure 6 Melt-pool area measurement from coaxial high-speed image analysis. (a) Single square scan experiments (Exp. 1-3). (b) Island scan experiments (Exp. 4-7).

Table 2. Average melt-pool area, standard deviation, and build time for Exp. 1-7.

Exp. #	1	2	3	4	5	6	7
Mean melt-pool area (μm^2)	*29900*	*33300*	*39700*	*41700*	*40000*	*42500*	*50300*
Standard deviation (μm^2)	*4400*	*4700*	*8900*	*4700*	*8300*	*8100*	*15900*
Build time (s)	*0.098*	*0.076*	*0.088*	*2.258*	*3.600*	*2.758*	*3.356*

Yeung, Ho; Lane, Brandon; Fox, Jason; Kim, Felix; Heigel, Jarred; Neira, Jorge.
"Continuous Laser Scan Strategy for Faster Build Speeds in Laser Powder Bed Fusion System."
Paper presented at The 28th Annual International Solid Freeform Fabrication Symposium, Austin, TX, United States. August 7, 2017 - August 10, 2017.

Based on the in-situ melt pool imaging, and ex-situ topography results, the following observations are made:

1. For single square scans (Exp. 1-3), continuous linear mode is the most time efficient, while continuous concentric mode has the largest mean melt-pool area. The increasing melt-pool area for Exp. 3 (Figure 6) should be due to the residual heat building up as concentric circles are getting smaller.
2. For multi-square island scans (Exp. 5-7), the trend is consistent with single square scans. The time for continuous linear mode (Exp. 6) improved by 23% over constant build speed mode (Exp. 5), and average melt-pool size for continuous concentric mode (Exp. 7) improved by 25% over constant build speed mode (Exp. 5).
3. Island scans (Exp. 4-7) yield a larger melt-pool area than their single square counterparts (Exp. 1-3). Longer scan time and more localized scan path helped build up the residual heat.
4. Comparing Exp. 4 with 5, island scan strategy does not seem to have any advantage for constant build speed mode, in terms of both building time and melt-pool size. The laser was turned on and off very frequently for Exp. 5 because of the shorter scan distance. That might have balanced out the localized heat effect of the island scan.
5. The standard deviations of the melt-pool areas are the smallest for Exp. 1 and 4. The power density is more consistent in constant build speed mode.

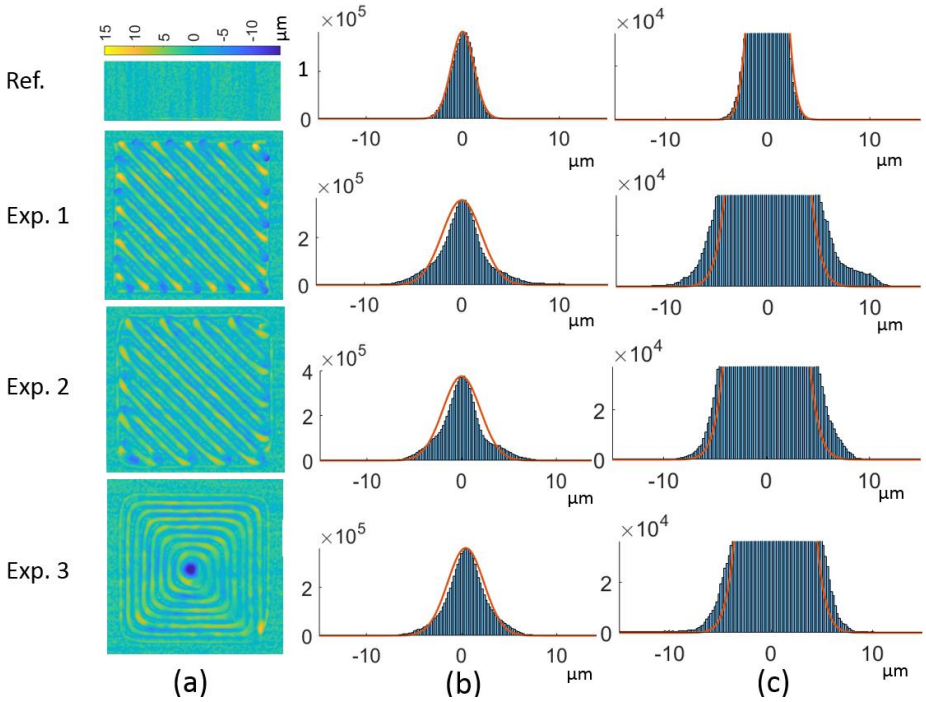


Figure 7. Surface profiles for Exp. 1-3. (a) Topographies. (b) Histogram of height data with normal distribution curve fitted. X-axis is the height in μm, Y-axis is number of points. (c) Enlarged view of (b).

Figure 7a shows the topographies for Exp. 1-3, and a reference region (an un-scanned area on the same metal plate). Figure 7b is the distribution of the height for each point on the topography. Figure 7c is an enlarged view of Figure 7b, and shows how the height data spread out around boundaries. The height profile for the reference fits very well into a normal distribution curve, which represents the surface

Yeung, Ho; Lane, Brandon; Fox, Jason; Kim, Felix; Heigel, Jarred; Neira, Jorge.
"Continuous Laser Scan Strategy for Faster Build Speeds in Laser Powder Bed Fusion System."
Paper presented at The 28th Annual International Solid Freeform Fabrication Symposium, Austin, TX, United States. August 7, 2017 - August 10, 2017.

roughness of the original metal plate. All experiments resulted in a 'rougher' surface as the height data spread wider (Figure 7b and Figure 7c), especially for Exp. 1 which goes beyond +/- 10 µm range. A rougher surface may end up a poorer part quality in a three-dimensional build as it prevents even distribution of powder, or it may even jam the recoater blade. To further visualize how different scan strategies affect the surface profile, the top left corners of each experiment were enlarged and compared in Figure 8.

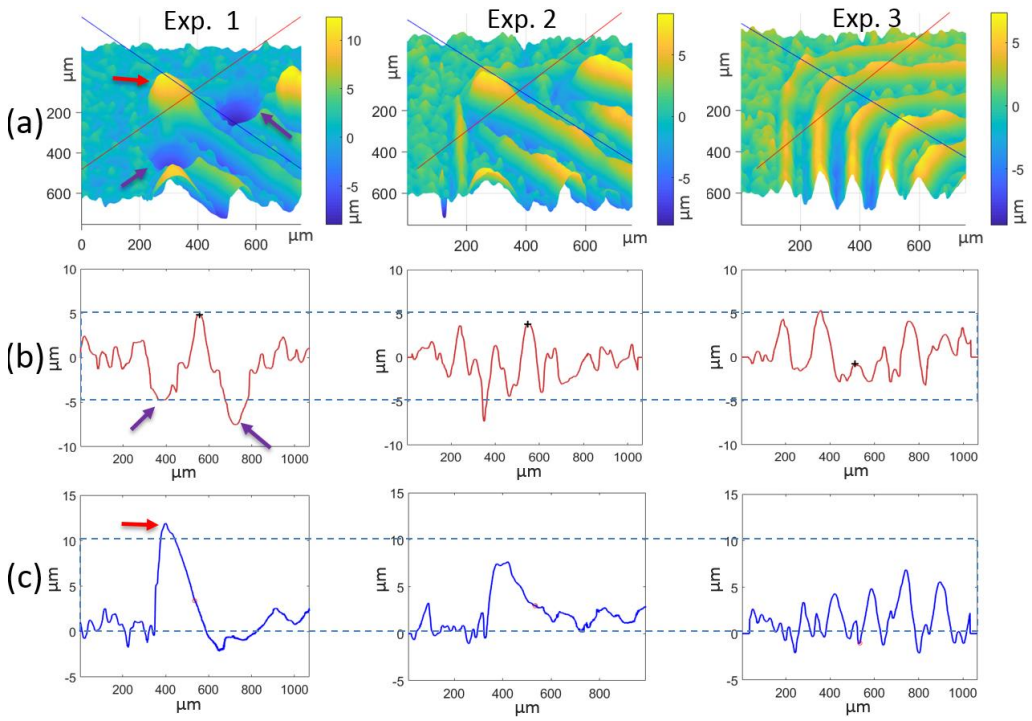


Figure 8. Surface profiles for Exp. 1-3. (a) Topographies. (b) Height profile along the red line. (c) height profile along the blue line.

For constant build speed mode (Exp. 1), the laser power was turned on at the beginning and turned off at the end for each scan line. That created bumps at the beginnings and holes at the ends, as indicated by the red and purple arrows in Figure 8a. Continuous linear mode (Exp. 2) reduces the laser on and off frequency, but the sharp turn between adjacent hatch lines creates complicated local thermodynamics. That still resulted in 'valleys' around turns, but at a smaller scale compared to Exp. 1. The continuous concentric mode (Exp. 3) has no sharp turns (except at the center), resulting in the most even surface, as shown in Figure 8b and Figure 8c.

Summary and Future work

There is an open field of research into laser scan strategy with the potential to reduce defects, control residual stress or microstructure, or improve the speed and efficiency of material consolidation. The NIST AMMT provides control and monitoring capabilities for such research. This paper demonstrated two advanced scan strategies implemented on AMMT – continuous linear and continuous concentric. The results show clear advantages over traditional (constant build speed) scan strategy for both efficiency and

Yeung, Ho; Lane, Brandon; Fox, Jason; Kim, Felix; Heigel, Jarred; Neira, Jorge.
"Continuous Laser Scan Strategy for Faster Build Speeds in Laser Powder Bed Fusion System."
Paper presented at The 28th Annual International Solid Freeform Fabrication Symposium, Austin, TX, United States. August 7, 2017 - August 10, 2017.

quality. The build time for the continuous linear strategy is 23% faster than the traditional strategy, and the continuous concentric strategy effectively reduced bumps and holes compared to the traditional strategy.

Constant laser power was used for all experiments in this work, although AMMT has the capability for adjusting laser power at 100 kHz. A proper tuning of laser power should be able to further improve the surface quality. It is also shown here that a continuous island scan produced a larger melt-pool, and the island scan is believed to be able to reduce residual stress as well. A study of multilayer powder builds based on varying power continuous concentric island scans should be very interesting.

References

[1] Mani M, Lane B, Donmez M A, Feng S, Moylan S and Fesperman R 2015 *Measurement science needs for real-time control of additive manufacturing powder bed fusion processes* (Gaithersburg, MD: National Institute of Standards and Technology)

[2] King W E, Barth H D, Castillo V M, Gallegos G F, Gibbs J W, Hahn D E, Kamath C and Rubenchik A M 2014 Observation of keyhole-mode laser melting in laser powder-bed fusion additive manufacturing *J. Mater. Process. Technol.* **214** 2915–25

[3] Thijs L, Verhaeghe F, Craeghs T, Humbeeck J V and Kruth J-P 2010 A study of the microstructural evolution during selective laser melting of Ti–6Al–4V *Acta Mater.* **58** 3303–12

[4] Yadroitsev I, Thivillon L, Bertrand P and Smurov I 2007 Strategy of manufacturing components with designed internal structure by selective laser melting of metallic powder *Appl. Surf. Sci.* **254** 980–3

[5] Khairallah S A, Anderson A T, Rubenchik A and King W E 2016 Laser powder-bed fusion additive manufacturing: Physics of complex melt flow and formation mechanisms of pores, spatter, and denudation zones *Acta Mater.* **108** 36–45

[6] Cheng B, Shrestha S and Chou K 2016 Stress and deformation evaluations of scanning strategy effect in selective laser melting *Addit. Manuf.* **12** 240–51

[7] Gockel J and Beuth J L 2013 Understanding Ti-6Al-4V microstructure control in additive manufacturing via process maps *Solid Freeform Fabrication Proceedings* Solid Freeform Fabrication Proceedings (Austin, TX) pp 666–74

[8] Mercelis P and Kruth J-P 2006 Residual stresses in selective laser sintering and selective laser melting *Rapid Prototyp. J.* **12** 254–65

[9] Yeung H, Neira J, Lane B, Fox J and Lopez F Laser Path Planning and Power Control Strategies for Powder Bed Fusion Systems *Solid Free. Fabr. 2016 Proc. 27th Annu. Int. Solid Free. Fabr. Symp.*

[10] Lane et al. Design, Developments, and Results from the NIST Additive Manufacturing Metrology Testbed (AMMT) *Solid Free. Fabr. 2016 Proc. 26th Annu. Int. Solid Free. Fabr. Symp.*

[11] Electronic Industries Association Interchangeable Variable Block Data Format for Positioning, Contouring, and Contouring/Positioning Numerically Controlled Machines *EIA Stand. EIA-274-Febr. 1979*

[12] Carter L N, Martin C, Withers P J and Attallah M M 2014 The influence of the laser scan strategy on grain structure and cracking behaviour in SLM powder-bed fabricated nickel superalloy *J. Alloys Compd.* **615** 338–47

[13] Kruth J P, Froyen L, Van Vaerenbergh J, Mercelis P, Rombouts M and Lauwers B 2004 Selective laser melting of iron-based powder *J. Mater. Process. Technol.* **149** 616–22

[14] Rombouts M, Kruth J-P, Froyen L and Mercelis P 2006 Fundamentals of selective laser melting of alloyed steel powders *CIRP Ann.-Manuf. Technol.* **55** 187–92

Yeung, Ho; Lane, Brandon; Fox, Jason; Kim, Felix; Heigel, Jarred; Neira, Jorge.
"Continuous Laser Scan Strategy for Faster Build Speeds in Laser Powder Bed Fusion System."
Paper presented at The 28th Annual International Solid Freeform Fabrication Symposium, Austin, TX, United States. August 7, 2017 - August 10, 2017.

Business Process Context for Message Standards

Nenad Ivezic[1,*], Miroslav Ljubicic[1], Marija Jankovic[1], Boonserm Kulvatunyou[1], Scott Nieman[2], and Garret Minakawa[3]

[1] National Institute of Standards and Technology, Gaithersburg, MD, USA
{nivezic,miroslav.ljubicic,marija.jankovic,serm}@nist.gov
[2]Land O'Lakes, Shoreview, MN, USA
stnieman@landolakes.com
[3]Oracle, Redwood City, CA, USA
garret.minakawa@oracle.com

Abstract. Despite unrelenting increase in complexity of message standards for enterprise systems integrations, there are no effective means to address this complexity issue in practice. We describe an effort to address the issue by advancing message standards development and use methods. The new effort relies on business process model life-cycle management, which is essential for context definition of message standards usage. Context is essential as it describes the intent for the message standards usage for a specific systems integration case. We report results of a preliminary assessment of the approach for an industry use case.

Keywords: systems integration, message standards, life-cycle management, business process model, context

1 Introduction

Efficient, practical, systems integration continues to be a great challenge for enterprises of all sizes, in great part because of the increasing complexities of message standards for the integration. The Open Applications Group, Inc. (OAGi) is one of the original consortia that standardize message-exchange standards [1]. Without a means to manage a shareable context specification, OAGi members have seen the message standards becoming complex and their management unwieldy.

Business processes are prime candidate to supply context specification for the messages involved in information exchanges. This has been recognized for many decades, starting with the activity modeling language IDEF0 where inputs and outputs capture the business data to be exchanged between activities [2]. The OAGi consortium has taken first steps to offer BPMN-based standards for business processes to provide precise context for message exchanges [3].

Recently, BPMN 2.0, with its BPMN.xsd representation and runtime execution capability, has accelerated the design, development, and implementation of message and process standards [3]. However, problems still exist in 1) consistency and interoperability between business process modeling tools, 2) adequacy of the content captured in the process model, and 3) process cataloging for reuse and adaptability.

Ivezic, Nenad; Ljubicic, Miroslav; Jankovic, Marija; Kulvatunyou, Boonserm; Nieman, Scott; Minakawa, Garret.
"Business Process Context for Message Standards."
Paper presented at Business Process Management, Barcelona, Spain. September 8, 2017 - September 15, 2017.

In this paper, we describe a new approach for business-process model life-cycle management (BPM LCM) to tackle these problems. The outcome of this approach will be a useful, shared, business process-based context definition for message standards development and use. Such a definition will allow enterprises to accelerate systems integration efforts.

2 What are Message Standards, anyway?

Message standards are data standards that define both the structure and the semantics of the message. Such standards govern information exchange among applications, services, and other actors. By doing so, message standards facilitate the systems integration. Effective information exchange, however, is hindered by the growth of individual message standards, in both size and complexity.

Presently, message standards, such as Open Applications Group Integration Specification (OAGIS), are growing complex for multiple reasons [1]. First, systems today are deployed on a variety of computing platforms, each using a different computer language. Second, these standards support a wide range of enterprise business processes and sectors including aerospace, automotive, chemical, and electronics that are subject to various quality criteria, regulations, and other factors. Third, new industrial integration use cases continue to expand definitions of message standards. Thus, critics often claim these standards are 'bloated.'

To use such "bloated" message standards today, most companies must perform manual, time-intensive adaptations of message standards to address systems integration requirements. These adaptations result in subsets or profiles of the standard for actual use in the context of each such systems integration case. To create this profile, implementers and business analysts must first determine which elements of the message standard are applicable to their integration use case. Then, they must manage and relate that part to edge application APIs. This results in time-intensive, error-prone, manual interpretations of standards. In addition, these efforts must be repeated every time a new computing platform is introduced.

As a strategy to address these problems, industry has shown interest in using life-cycle management (LCM) methods to advance message standards. LCM methods are processes for the development, use, adaptation, operation, and maintenance of a standard. These methods are expected to cover all phases of the standard from the requirements gathering to the end of life. Examples of OAGIS message standards, and their associated LCM methods, can be found in [1].

The current message-standards LCM methods, however, lack the ability to manage the growing complexity of message standards. The reasons are 1) they treat integration use cases independently and 2) they provide only for additions of required data elements to standard message definitions. In other words, these methods do not capture the underlying business processes that drive the integration in the first place; nor, do they identify the data elements that are shared as part of that integration. Consequently, there is no consistent, shareable definition of the intended integration uses of any message standard – that is, definition of context.

Ivezic, Nenad; Ljubicic, Miroslav; Jankovic, Marija; Kulvatunyou, Boonserm; Nieman, Scott; Minakawa, Garret.
"Business Process Context for Message Standards."
Paper presented at Business Process Management, Barcelona, Spain. September 8, 2017 - September 15, 2017.

Such a definition could inform and specify those intended uses including the necessary adaptations and refinements of message standard across different integration situations. To do so, the definition must provide usage information that includes (1) customized or profiled message standard, (2) intent for the customized message standard, and (3) accumulation of data at each step of the business process used to customize the message standard.

Our work addresses the absence of the means to provide and manage usage information of messaging standards. To understand our approach, consider the current (As-Is) state of message-standards life-cycle management (MS LCM) in Fig. 1 and the envisioned (To-Be) state of MS LCM in Fig. 2. There are three areas where we seek advancements in MS LCM (as indicated in the figures), which currently have very limited tool support:

1. *Integration Requirements Definition* where today Business Process Analysts inefficiently and in an ad-hock manner specify the requirements in natural language (typically using non-standard business process models created in Visio or Powerpoint) and based on a target business process to be supported.
2. *Message Standards Adaptation* where today Software Developers inefficiently work with the integration requirements provided in natural language form to identify and adapt (e.g., prune and extend) standard messages for specific application schemes present in the integration case. The developers also review application APIs to identify required fields that may have been missed, and may refactor message standard profile to include these fields.
3. *Profile Message Generation* where presently Software Developers engage in cost-inefficient transformations of one implementation language-specific profile message definition into another.

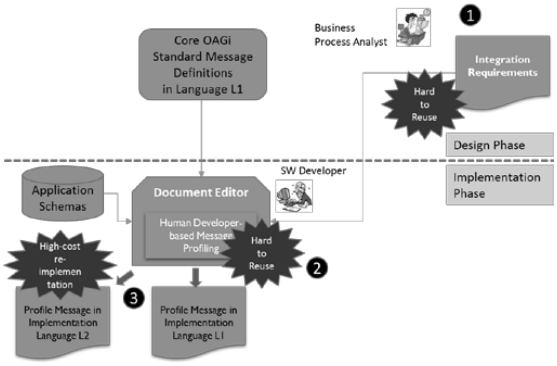


Fig. 1. As-is state of Message Standards Life-Cycle Management

Our focus in this paper is on the first area where a new business process model life-cycle management (BPM LCM) approach is introduced for message standards

Ivezic, Nenad; Ljubicic, Miroslav; Jankovic, Marija; Kulvatunyou, Boonserm; Nieman, Scott; Minakawa, Garret.
"Business Process Context for Message Standards."
Paper presented at Business Process Management, Barcelona, Spain. September 8, 2017 - September 15, 2017.

context definition management. The approach allows greater reuse and automation in the Integration Requirements Definition area, and in the other two areas of interest.

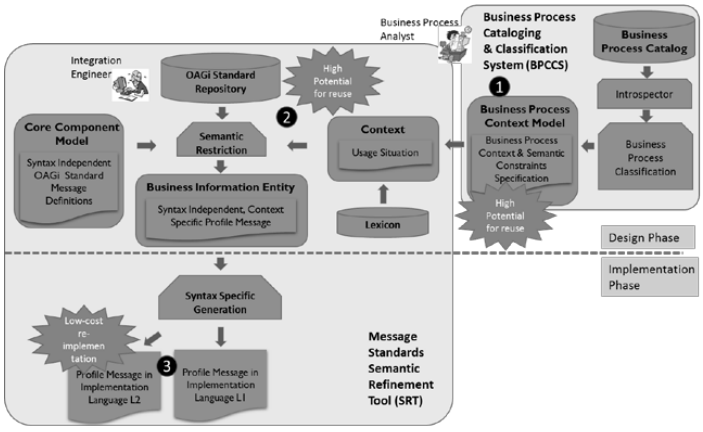


Fig. 2. To-be state of Message Standards Life-Cycle Management

3 Tools to migrate from the As-Is to the To-Be state

Our research, which was done with OAGi, has led to the development of two tools: 1) Message Standards Semantic Refinement Tool (MSSRT), which improves the message standards LCM process; and 2) Business Process Cataloging and Classification System BPCCS) to link process models to a usage meta-model, enabling Business Process Models Life-Cycle Management (BPM LCM).

3.1 Overall Architecture

Fig. 2 shows responsibilities of the two tools – BPCCS and MSSRT – in the To-Be state of MS LCM. BPCCS is being developed as part of a BPM LCM method to manage the life-cycle of both reference and context-specific OAGIS business-process models. A core focus of the BPCCS tool is to provide shared definitions of required concepts and terms for business processes that span across an enterprise and its multi-tier supplier network. The BPCCS tool performs three major functions. The first is to create and manage the Context Model. The second is to provide to the Business Process Analyst a user interface by which the context model is specified along with additional semantic constraints on the process model. The third is to communicate this usage information to the MSSRT tool.

The MSSRT tool will apply this usage information to the syntax-independent, standard, message definition to create a context-specific profile message. MSSRT also transforms a syntax-independent form of the profile message into an implementation-specific profile message. In providing these capabilities, the

Ivezic, Nenad; Ljubicic, Miroslav; Jankovic, Marija; Kulvatunyou, Boonserm; Nieman, Scott; Minakawa, Garret.
"Business Process Context for Message Standards."
Paper presented at Business Process Management, Barcelona, Spain. September 8, 2017 - September 15, 2017.

MSSRT tool will enable business-process-model discovery and reuse. Towards that goal, we are planning to use introspection functionality (allowing discovery of the model properties at runtime) to harvest business- process-model information for context definition.

MSSRT is intended to aid the systems integrators and users in generating and cataloging the message-standard profiling information using a new, CCS-compliant OAGIS meta-model [4]. This tool is used as a web-based, application-software environment for the life-cycle management of message standards. In parallel, the tool will be utilized to experiment with new methods to create, maintain, and use message standards. The software tool includes collaborative, multi-tenant methods and meta-models for life-cycle management of message standards. Those methods and models can be shared with the community and can facilitate the creations of extensions made by the community to be submitted into the standard. It provides core functionalities that can be extended and commercialized by industry. Among the functionalities being explored are those to allow us to deal with natural language issues, such as term matching and synonym handling (indicated in Fig. 2 as Lexicon).

The following sections focus on our systems engineering approach in developing BPCCS, the current state of the development, and initial tool assessment results.

3.2 Business Process Cataloging and Classification System (BPCCS)

Requirements Gathering. We collected use cases to gather requirements and identify activities to support using BPM LCM functions of the BPCCS. For example, one use case is an end-to-end, product-procurement scenario with the goal to support BPM discovery and reuse. The case starts with the customer issuing a purchase order and ends with the customer receiving the goods, the shipment notification, and, the as-built inspection information. The components of BPMs are classified using a reference classification framework APQC PCF [5]. Multiple. context-classification schemes are managed in the BPCCS to help catalog these BPMs, including ISO 10314 classification [6], ebXML-adopted Porter classification [7], and OAGIS functional classification [5]. When outsourcing the enterprise activities associated with parts of a BPM, the BPCCS provides BP-based context information to help discover available manufacturing services. Once services are discovered, the original BPM may be modified. The original BPM is kept for traceability purposes, where the APQC PCF is employed to allow cross-industry reference.

Requirements Analysis. Fig. 3 illustrates the two main functional parts of our approach to support the required BPM LCM functions: Classification Schemes (on the right) and Catalog (on the left).

The Classification Schemes allow the BPCCS users to classify their process models (and their parts) using multiple contextual dimensions, thereby defining a context for the intended use of the model. Several contextual dimensions have been proposed previously for inclusion into systems like BPCCS. Those dimensions include industry, product, business process, role, and function, among others [8]. The

Ivezic, Nenad; Ljubicic, Miroslav; Jankovic, Marija; Kulvatunyou, Boonserm; Nieman, Scott; Minakawa, Garret.
"Business Process Context for Message Standards."
Paper presented at Business Process Management, Barcelona, Spain. September 8, 2017 - September 15, 2017.

assumption is that it should be possible to find a process model with the needed (or similar) semantics by providing the context in which the model will be used.

The BPCCS Catalog stores and inspects business process models to extract relevant metadata. Both reference and specific models are stored and described by the context in which they are derived using Classification Schemes' context dimensions. Derived process models are regarded as variants of the reference process model. Variants are needed because the context of a specific model can differ from the context of the original reference model, as illustrated in Fig. 3. Reference- model elements are used as shared terminology to which elements of specific process models are mapped and thus semantically aligned.

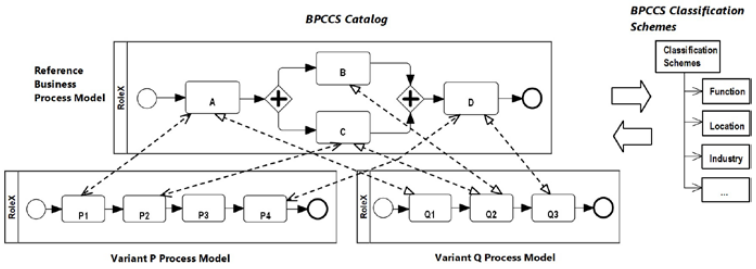


Fig. 3. Overview of Approach

Conceptual Design. Following the requirements analysis, we identified the needed BPCCS capabilities, chief among which was the ability to capture precisely context and its semantics for BPM LCM. We analyzed research results in four related areas regarding these capabilities [9]: Modeling BP Variability; Modeling BP Context; BP Catalog Approaches; and Industry Approaches. Our analysis showed that effective support for the needed capabilities still does not exist. (Please see [9] for details of the analysis.) To address this situation, we proposed a BPCCS metamodel that builds on the ebRIM specification [10].

BPCCS allows its users to classify their process models, and parts of the models, using multiple contextual dimensions. The dimensions can be grouped by their purposes and organized into 'aspects', represented by the ContextAspect element. Currently, we are leveraging Zachman's framework [11] and its 5WH maxims[1] to organize context dimensions into context aspects. For example, geographical-location and organization-unit context dimensions is related to the 'Where maxim'; while, the industry context dimension is related to the 'What maxim.' The previous Introspection functionality can help pre-populate the results of the 5WH maxims.

Fig. 4 illustrates the metamodel concepts on a generic example. Here, the context of a business process is described by associating the process (the far right of the Fig. 4) with various (multiple) classification nodes (OY, US, 44, 10279, etc.), belonging to different classification schemes (ISO 3166, NAICS, etc.) used to describe different

[1] Who, What, Where, When, Why, and How

dimensions of the context (Role, Geo Location, etc.) through giving the answers to different context aspects (Why?, When?, etc.).

To define the range of values for a context dimension, different taxonomies or controlled vocabularies can be used. Beside classification schemes, stored catalog objects of a particular type can also be used as values of context dimensions.

Finally, the context of a particular catalog object is defined as the set of all the associations the object has. Included among the associations are classification nodes of appropriate schemes and other catalog objects of appropriate types that can be used for particular context dimensions.

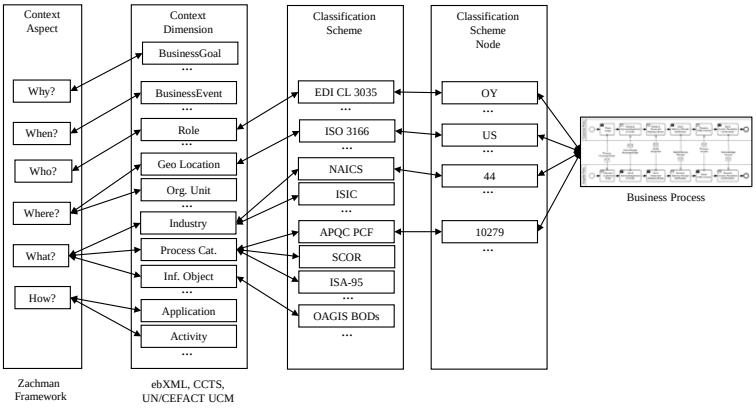


Fig. 4. An Illustration of Context for a Business Process Model

Verification. We have developed proof-of-concept prototype based on the BPCCS metamodel. Also, we collected various business processes together with their models and context definitions, which we used to populate the prototype implementation. Most of these processes were obtained from members of the OAGi consortium where integration use cases are collected for enhanced reuse of information and communication artifacts. In this example, the goal is to describe business processes and their components for their subsequent retrieval and reuse.

We focus on *Retrieve Electronic Control Unit (ECU) Information* sub-process which encapsulates activities for exchanging information about the product and its parts (in this case, ECU) between various systems [12]. The goal is to design this sub-process by reusing an existing process model.

First, we needed to search the collection of populated, business-process models for the ones applicable to the context of the *Retrieve ECU Information* sub-process. To use this context, values for all context aspects should be provided. As suggested before, this can be done by answering Zachman's interrogatives. Context for *Retrieve ECU Information* sub-process, defined in this way, is given in Fig. 5.

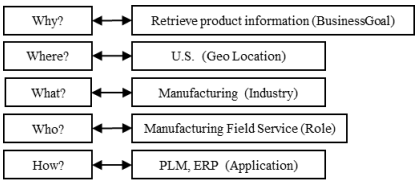


Fig. 5. Example Context for the *Retrieve ECU Information* Sub-process

Once the context is identified, we can browse BPCCS to find applicable business processes. For this purpose, a complete or partial context from Fig. 5 can be used. For example, we can start by searching processes that achieve the goal (Retrieve product information). This search yielded three business processes models – three different BP variants of the Exchange Product Information goal.

Although the outcome of all variants is the same – exchanging full product information – each variant uses a different sequence of activities and different message exchanges. Messages are defined using OAGIS Business Object Documents (BODs) specification [1]. BODs are defined using a verb-noun structure, where verb defines the desired action that should be applied to the exchanged business information, which is represented by a noun. For example, ProcessBOM BOD (verb: Process; noun: BOM – Bill of Materials) specifies that the receiving system shall execute a certain process on the BOM contained in the BOD.

Variant 1 uses BODs based on a single noun – ExhaustiveBOM noun, which defines BOM for full product information, including structure and child item details. Variant 2, instead of exchanging full product information with a single message, uses a BOD based on StructureBOM noun to exchange structure of BOM first. Variant 3 is same as Variant 2, but with one subtle difference. After receiving the structure of the product, the user decides which detailed parts information should be retrieved. This was not possible in Variant 2, where details for all product's parts were exchanged by default.

These differences between variants are related to their different, although similar, contexts, as shown in Table 1. Columns represent variants, while rows define context dimension/aspect combinations. Table 1 also shows used classifications.

Analyzing contexts of the business process variants provide information for determining which variant should be reused. For example, there is a difference in the Industry context dimension, where Variant 1 is designed for Wireless Telecommunications Carriers, and Variant 2 and Variant 3 are designed for Electronic Computer Manufacturing. This context difference had a crucial impact on variant design and is correlated with the messages (BODs) used in the variant. Namely, telecommunication products and services do not have overly complex structure; and, it is computationally feasible to exchange their full information using a single message. Hence, Variant 1 uses BODs based on ExhaustiveBOM noun. However, this cannot be expected for manufacturing domain as products are typically much more complex. Thus, Variant 2 and Variant 3 separate product

information using BODs based on StructureBOM and ItemMaster nouns to make potentially very large product information exchange feasible.

Table 1. Comparison of three contexts for retrieved business process variants.

	Exch. product info. v1	**Exch. product info. v2**	**Exch. product info. v3**
Business Goal (Why?)	*Class. Scheme:* Custom *Node/Value:* Retrieve product information, Synchronize product information	*Class. Scheme:* Custom *Node/Value:* Retrieve product information, Synchronize product information	*Class. Scheme:* Custom *Node/Value:* Retrieve product information
Process Category (What?)	*Class. Scheme:* APQC Cross-industry PCF (7.0.4) *Node/Value:* Manage product and service master data	*Class. Scheme:* APQC Cross-industry PCF (7.0.4) *Node/Value:* Manage product and service master data	*Class. Scheme:* APQC Cross-industry PCF (7.0.4) *Node/Value:* Manage product and service master data
Industry (What?)	*Class. Scheme:* NAICS 2012 *Node/Value:* 517210 Wireless Telecommunications Carriers	*Class. Scheme:* NAICS 2012 *Node/Value:* 334111 Electronic Computer Manufacturing	*Class. Scheme:* NAICS 2012 *Node/Value:* 334111 Electronic Computer Manufacturing
App. (How?)	*Class. Scheme:* Custom *Node/Value:* CRM, ERP, API	*Class. Scheme:* Custom *Node/Value:* CRM, ERP, API	*Class. Scheme:* Custom *Node/Value:* CRM, ERP, API
Role (Who?)	*Class. Scheme:* Custom *Node/Value:* Customer Relationship Management, Integration Specialist, Order Management	*Class. Scheme:* Custom *Node/Value:* Customer Relationship Management, Integration Specialist, Order Management	*Class. Scheme:* Custom *Node/Value:* Customer Relationship Management, Integration Specialist, Order Management
Location (Where?)	*Class. Scheme:* ISO 3166 *Node/Value:* US	*Class. Scheme:* ISO 3166 *Node/Value:* US	*Class. Scheme:* ISO 3166 *Node/Value:* US

4 Discussion and Next Steps

Our assessment showed that context specification enabled by the BPCCS metamodel supports desired behaviors. First, it was possible to specify the goal context intuitively using Zachman's 5WH interrogatives. Second, it was possible to search for business processes applicable to a given context, as described by context dimensions. Third, manual comparison of business processes by their context was supported. Fourth, process-model designs could be analyzed in correlation with their contexts. Fifth, it was possible to find, and manually reuse, the most appropriate business process, together with message profiles (i.e., BOD subsets).

Other types of assessments are planned including support for greater automation in business process model reuse and refinement, which requires further development of context management apparatus. This is planned to be done using semantic technologies, developed in parallel with a needed ontological basis, for explicit and shared conceptualization of the context elements. For the automation to be realized, BPMN [3] conformance testing of process modeling tool is anticipated, which is necessary for the business process models introspection functionality.

Finally, we are planning for user and organizational adoption of MSSRT and BPCCS tools, which is a particularly challenging in the light of likely disruptions to the current practices of message standards development and use.

Ivezic, Nenad; Ljubicic, Miroslav; Jankovic, Marija; Kulvatunyou, Boonserm; Nieman, Scott; Minakawa, Garret.
"Business Process Context for Message Standards."
Paper presented at Business Process Management, Barcelona, Spain. September 8, 2017 - September 15, 2017.

5 Conclusion

The paper presents a new approach to manage business process-based context definition that describes the intent for usage of message standards. Central to the approach is business process model life-cycle management capability. A preliminary assessment shows that the approach provides desired support to the end user in search, comparison, analysis, and reuse of business process models. Next steps include integration of the approach within the overall message standards life-cycle management support, further validation of the approach, and work on its adoption within standards development organizations.

Disclaimer

Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.

References

1. Open Applications Group Integration Specification (OAGIS) . http://www.oagi.org. Accessed 12 Jul 2017.
2. NIST (1993) Draft Federal Information Processing Standards, Standard for Integration Definition for Function Modeling (IDEF0).
3. BPMN (2011) Business Process Model and Notation (BPMN), Version 2.0.
4. ISO 15000-5:2014 Electronic Business Extensible Markup Language (ebXML) - Part 5: Core Components Specification (CCS). https://www.iso.org/standard/61433.html. Accessed 12 Jul 2017
5. APQC Process Classification Framework. https://www.apqc.org/. Accessed 12 Jul 2017
6. ISO/TR 10314-1:1990 Industrial automation - Shop floor production - Part 1: Reference model for standardization and a methodology for identification of requirements. https://www.iso.org/standard/18360.html. Accessed 12 Jul 2017
7. ebXML (2001) ebXML Catalog of Common Business Processes v1. 0. UN/CEFACT and OASIS
8. ebXML (2001) The role of context in the re-usability of Core Components and Business Processes. UN/CEFACT and OASIS
9. Ljubicic M, Ivezic N, Kulvatunyou B, et al. (2017) Business Process Model Life-Cycle Management in Cloud Manufacturing. Proceedings of the ASME 2017 International Manufacturing Science and Engineering Conference
10. ebXML RIM (2012) OASIS ebXML RegRep Version 4.0 Part 1: Registry Information Model (ebRIM). OASIS
11. Zachman J (2002) The Zachman Framework for Enterprise Architecture.
12. Kulvatunyou, B., Ivezic, N., and Srinivasan V. "On architecting and composing engineering information services to enable smart manufacturing." Journal of computing and information science in engineering 16.3 (2016).

Production System Identification with Genetic Programming

Peter Denno[a,1], Charles Dickerson[b] and Jenny Harding[b]
[a]National Institute of Standards and Technology, US
[b]Loughborough University

Abstract. Modern system-identification methodologies use artificial neural nets, integer linear programming, genetic algorithms, and swarm intelligence to discover system models. Pairing genetic programming, a variation of genetic algorithms, with Petri nets seems to offer an attractive, alternative means to discover system behaviour and structure. Yet to date, very little work has examined this pairing of technologies. Petri nets provide a grey-box model of the system, which is useful for verifying system behaviour and interpreting the meaning of operational data. Genetic programming promises a simple yet robust tool to search the space of candidate systems. Genetic programming is inherently highly parallel. This paper describes early experiences with genetic programming of Petri nets to discover the best interpretation of operational data. The systems studied are serial production lines with buffers.

Keywords. System identification, Petri nets, genetic programming, smart manufacturing

1. Introduction

The ability to generate models of manufacturing systems from data is becoming increasingly useful. In earlier generations of manufacturing systems, a model of the system as a discrete-event system could be developed through inspection of the system's controller software. But, the machine-learning technology that is playing an increasing role in manufacturing control today [1] does not provide a similar presentation of system structure and mechanism. For this reason, verifying system behaviour and safety are becoming more, not less, challenging. Because of this, systems identification methodologies that suggest system structure (grey box models) are becoming increasingly useful. [2]

This paper presents a mostly-automated methodology to produce Petri net (PN) models of systems from historical, operational data describing system inputs and outputs. PNs used in this capacity provide grey-box, system models of the system that suggest the system's structure. Such structurally accurate PN models traditionally have been used for several purposes including verifying system safety, assessing system performance, and detecting deadlocks. This paper describes a method to match differing interpretations of system characterization with the appropriate Petri net structure. For example, multiple interpretations of the notions of blocking and starvation are prevalent in productions-systems engineering. Each interpretation associates accurately with some

[1] Corresponding author

Denno, Peter; Kulkarni, Amogh; Balasubramanian, Daniel; Karsai, Gabor.
"Production System Identification with Genetic Programming."
Paper presented at 2018 IEEE International Conference on Industrial Technology (ICIT), Lyon, France. February 19, 2018 - February 22, 2018.

Petri net structure. The goal of this work is to determine which interpretation most closely matches the given operational data.

The paper describes preliminary work towards that goal. It uses a case study involving serial production lines. [3] The system-identification algorithm used in this study is provided with two types of inputs: product mixes and machine capacities. The algorithm then generates the corresponding outputs in the form of four, steady-state, performance measures: buffer occupancy, probability of blocking, starvation and throughput. The algorithm applies genetic programming (GP), an evolutionary programming technique, to generalized stochastic Petri nets (GSPNs) to discover a Petri net that best fits the system input-output relations.

The contribution of this paper is a report of early experience using GP with GSPNs. The authors are aware of only one paper of a similar nature [4] and the case study of that paper concerns biological processes, not manufacturing processes, and discrete Petri nets, not timed Petri nets.

Section 2 of the paper discusses related work and the fundamental concepts of Petri nets and genetic programming. Section 3 describes our methodology. Section 4 concludes the work.

2. Related Work

The semantics of GSPNs are described with the help of Figure 1, which depicts a GSPN modelling a system consisting of two machines and a buffer of capacity one between them. The figure, as is typical of Petri net notation, uses 1) solid bars to represent immediate transitions and 2) hollow bars to represent exponentially timed transitions [5] 3) hollow circles to represent places, which can be populated with tokens, and 4) solid circles to represent *m1-blocked*, *buffer* and *m2-busy*. States of the system are described by the quantities of tokens in places. Figure 1 depicts a GSPN describing a system consisting of two machines that process jobs using times from exponential distributions. Between the first machine and the second is a buffer with capacity for one job. In the net on the left side of the figure, both machines *m1* and *m2* are busy; on the right side, machine *m1* is blocked.

GSPNs allow two kinds of arcs between transitions and places, both used in the figure. Activating arcs have an arrow head. Inhibitor arcs have a hollow circle head. A multiplicity is associated with arcs. The execution semantics of GSPNs is as follows. A transition fires when 1) all the input places (places at the tail of an arc) have quantities of tokens equal to or greater than the multiplicity of the arc into the transition, and no input places have quantities of tokens equal to or exceeding the multiplicity of an inhibitor arc into the transition. When the transition fires, one or more tokens, equal to the multiplicity of the input activator arcs, are removed from the input places. Token are added to the output places based on the multiplicity of the output activator arcs. The multiplicity of all arcs in the figure is 1.

Petri nets are one among several representations applied in system identification. Fu and Li [2] survey modern methods of system identification including neural nets, fuzzy logic, genetic algorithms, and swarm intelligence. Though the authors do not specifically discuss genetic programming, many of their comments regarding genetic algorithms may also apply to genetic programming. With respect using genetic algorithms, they note two benefits: quick convergence and path independence; and, one potential drawback: premature convergence to local optima.

Denno, Peter; Kulkarni, Amogh; Balasubramanian, Daniel; Karsai, Gabor.
"Production System Identification with Genetic Programming."
Paper presented at 2018 IEEE International Conference on Industrial Technology (ICIT), Lyon, France. February 19, 2018 - February 22, 2018.

Nobile et al. [4] describe a methodology for evolving Petri nets for purposes such as system identification. Their test case is a metabolic process, not manufacturing. In their methodology, places and transitions are partitioned into visible and hidden subsets. Each visible place and transition is permanently associated with a domain quantity. Hidden transitions and places can be subject to removal by the evolutionary operations of crossover and mutation. Nobile et al. do not specifically address timed Petri nets, and therefore, do not suggest a means of setting transition rates. The use of GSPN in the present work necessitates a different set of genetic operators than those described by Nobile et al.

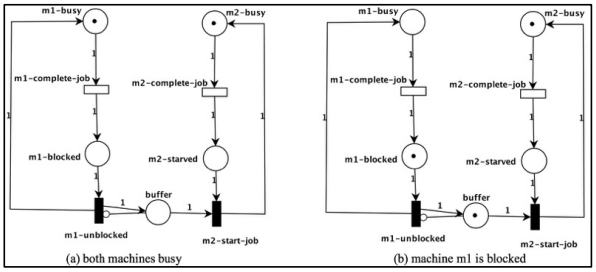


Figure 1. Two machines with a buffer for one part, block-after-service buffering convention.

Several works use Petri nets or genetic algorithms for systems identification. Tiacci [6] couples a discrete event simulator with a genetic algorithm approach to solve an assembly line balancing problem. Dotoli et al. [7] describe a method of system identification using Petri nets and integer linear programming. The work concerns the identification of discrete event systems as untimed Petri nets. In that work, the process is viewed as on-line, in the sense that it waits for events to occur and updates the Petri net after these occurrences. Basile et al. [8] also apply mixed integer linear programming. The Petri nets of this work are deterministic timed, not stochastic. The goal here is to provide a model that matches behaviour as discrete events. Rozinat [9] et al. uses process mining of event logs to create a simulation model of business processes as a coloured Petri Net. The work takes a broad perspective, involving identification of roles and merging of perspectives. Cabasino [10], a PhD thesis, describes an integer programming method of system identification and fault detection using unlabelled Petri nets. The goal of El Medhi et al. [11] is closest to the goal of the present paper. El Medhi describes an identification process for deterministic and stochastic (exponential) Petri net. Such nets can accurately represent queueing systems and machine reliability. The paper describes an integer linear programming method to synthesize a PN from measureable and non-measurable PN states.

3. Genetic Programming of Stochastic Petri Nets

Genetic algorithms are evolutionary algorithms. In a genetic algorithm, a population (sometimes called a generation) of individual solutions are scored for fitness relative to some objective. Those individuals scoring well are more likely to be promoted to the subsequent generation. Those solutions not selected are discarded. Among the promoted individuals, some are subject to modification by the application of two genetic operators:

Denno, Peter; Kulkarni, Amogh; Balasubramanian, Daniel; Karsai, Gabor.
"Production System Identification with Genetic Programming."
Paper presented at 2018 IEEE International Conference on Industrial Technology (ICIT), Lyon, France. February 19, 2018 - February 22, 2018.

mutation and crossover. The mutation operator modifies a single, selected individual; The crossover operator swaps elements of two individuals.

Genetic programming [12] is a form of genetic algorithm in which the individuals describe programs, typically represented as trees, and the operators are algebraic. The fitness function scores the ability of such a program to match the input/output relationships provided by training data. In applications where the best match can be represented as a mathematical function, $f: \mathcal{R} \rightarrow \mathcal{R}$, the problem closely resembles regression analysis. Indeed, the problem is a method of system identification called *symbolic regression* (SR). [13] Moreover, it can be viewed as a grey-box method if the discovered system provides a structure corresponding to a physical reality.

Our "programs" are Petri nets. [4] The elements of the programs are, of course, composed of the elements of whatever kind of PN has been selected. In this paper, we are using GSPNs because of their ability to model manufacturing systems. The manufacturing system we intend to examine in this paper concerns a 2-machine, serial production system with exponential service times and a one-place buffer between the machines. System identification in this context involves finding a Petri net that best matches the input/output relationships of the intended system. The chosen outputs are three steady-state properties: buffer occupancy, blocking of machine m1, and starvation of machine m2. These properties were used in the comparing the predicted values against the values of the intended system.

The design space of GSPNs has discrete and continuous dimensions. The discrete dimension concerns the PN's network topology. The continuous dimension concerns the real-valued rates of timed transitions and real-valued weights of immediate transitions. As is typical of generative design problems like SR, some strategy is needed to cope with the discrete/continuous dichotomy. In the design of other SR systems, that strategy uses 1) genetic programming to specify the discrete terms and 2) linear least-squares fitting to determine the optimal values for the continuous elements. Those values are the optimal coefficients of the terms in a linear function that minimizes prediction error. [13]

We have implemented a similar strategy in the design the SR system for our manufacturing example. There, the continuous terms, which are the service times and transition rates, are defined to be exponential with a mean of 1. The discrete term, the topology, evolves through the genetic program. The mutation operators used in that program are responsible for producing variations in the population.

Similar to the one in Nobile et al., [4] our GSPN design method makes a distinction between visible, and hidden, places and transitions. A *visible* place or transition is one for which system observations are provided. For this reason, the visible elements must appear in every PN. A *hidden* place or transition is one not directly associated with a system observation. Therefore, hidden elements need not be included in the PN. Given the visible/hidden dichotomy, the design of our GSPN focuses on visible elements only.

The mutation operators, as well as their arguments, in our GSPN are described in Table 1. In the application of the operators, a modified individual is promoted to the next generation only if it is found to be feasible. (Its reachability graph is calculated to determine this.) If the individual resulting from the mutation fails the feasibility test, a new set of random elements is selected and the individual is retested. If the selected mutation is not possible anywhere in the individual's structure, the individual is not promoted as a mutated form.

Table 1. Mutation operations used in the case study.

Operator	Action

add_place (t1, t2)	Two random transitions, t1, t2 (timed or immediate) are selected and a hidden place, p, and two arcs a1, and a2, are created. a1 is directed from t1 to p. a2 is directed from p to t2. The new arcs have multiplicity 1.
add_token(p)	A token is added to randomly selected place p (visible or hidden).
add_trans(p1, p2)	Two distinct places, p1 and p2 (visible or hidden) are chosen and a timed transition, t, and two arcs, a1 and a2 are created. a1 is directed from p1 to t. a2 is directed from t to p2. The new transition has rate=1.0.
add_arc(p,t)	A random place and transition (without regard to hidden/visible) are selected and randomly, either arc a is directed from p to t or t to p. The multiplicity of the arc is 1.
add_inhibitor(p,t)	A random place, p and transition, t is selected and an inhibitor arc, i, is created and directed from p to t.
remove_place(p)	A random hidden place is selected. It and all arcs to and from it are removed.
remove_token(p)	A place containing at least one token is randomly selected and its token count reduced by 1.
remove_trans(t)	A random hidden transition, t, is selected. It and all arcs to and from it are removed.
remove_arc(a)	A random arc, a, is removed.
remove_inhibitor(i)	A random inhibitor arc, i, is removed.
swap_places(v1, v2)	Two visible places are selected randomly. v1 is assigned the in-coming and out-going arcs of v2 and vice versa.

All selection actions used in the algorithm use tournament selection. Tournament selection involves choosing n individuals randomly from the population and then selecting the best among those n. Thus, when n is high, weak (low scoring) individuals are less likely to be selected. Selection is based on scoring individuals, which involves calculating the steady-state properties of the Petri net and comparing them to the target data. Since the individuals are stochastic Petri nets, this involves 1) forming the infinitesimal generator matrix **Q** for the Markov chain isomorphic to the Petri net and 2) solving the linear system:

$$\begin{aligned} \boldsymbol{\eta Q} &= \mathbf{0} \\ \boldsymbol{\eta 1}^T &= 1 \end{aligned} \tag{1}$$

where $\boldsymbol{\eta}$ is the steady-state distribution vector for the states of the Petri net. [5] For the case study, the dimension of **Q** was typically around 50 for most individuals but ranged as high as 700.

An initial population is created by producing a ring topology-PN called the Eden individual. The Eden individual contains all the visible places and transitions plus additional hidden places or transitions as needed to ensure that there are as many places as transitions. With equal numbers of places and transitions, the ring topology PN is produced by adding arcs between alternating places and transitions and closing the graph by connecting the last element used to the first. The Eden individual is repeatedly subjected to the genetic operators to create all of the individuals in the initial population.

In the case study, the visible places were labelled *m1-busy*, *m2-busy*, *buffer*, *m1-blocked,* and *m2-starved* as depicted in Figure 1. The system under study follows the usual conventions for analysis of serial production systems [3]: the first machine cannot starve and the last machine cannot block. The training data corresponded to machines with exponential service time. ($\lambda = 1.0$ for both machines.) The case study system uses block-after-service blocking convention.

Experience with the case study problem suggests that the algorithm essentially works, but that more effort will be needed to avoid convergence to local optima. As Table 2 shows, the algorithm tended to find a local optimum quickly and stick with it while the

Denno, Peter; Kulkarni, Amogh; Balasubramanian, Daniel; Karsai, Gabor.
"Production System Identification with Genetic Programming."
Paper presented at 2018 IEEE International Conference on Industrial Technology (ICIT), Lyon, France. February 19, 2018 - February 22, 2018.

median individual slowly improved. This was the case even when tournament selection pressure was low. The population size of the case study is 100 individuals. Bloat (the tendency in GP for individuals to become increasingly complex with little performance gain) was not a problem.

Table 2: Preliminary results from the case study

Generation	Error of Best Individual (total absolute error)	Error of Median Individual (total absolute error)
0	0.684	> 100.
1	0.684	1.80
2	0.494	1.33
8	0.470	0.855
9	0.333	0.720
10	0.333	0.512
12	0.333	0.467

4. Conclusion

The paper described early experience with a system-identification methodology that applies genetic programming to Petri nets representing discrete-event systems. The goal of the work is to discover Petri nets that best interpret the operational data. Genetic programming promises an effective method that easily parallelizes this problem. Early results suggest that the methodology is sound but that more work will be required to make it effective. We are currently undertaking that work.

References

[1] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine learning in manufacturing: advantages, challenges, and applications," *Prod. Manuf. Res.*, 2017.

[2] L. Fu and P. Li, "The Research Survey of System Identification Method," in *2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics*, 2013, pp. 397–401.

[3] J. Li and S. M. Meerkov, *Production System Engineering*. Springer Science+Business Media, 2009.

[4] M. S. Nobile, D. Besozzi, P. Cazzaniga, and G. Mauri, "The foundation of Evolutionary Petri nets," *CEUR Workshop Proc.*, vol. 988, pp. 60–74, 2013.

[5] M. A. Marsan, G. Balbo, G. Conte, and G. Franceschinis, "Modelling with generalised stochastic petri nets," *System*, p. 299, 1994.

[6] L. Tiacci, "Coupling a genetic algorithm approach and a discrete event simulator to design mixed-model un-paced assembly lines with parallel workstations and stochastic task times," *Int. J. Prod. Econ.*, vol. 159, pp. 319–333, 2015.

[7] M. Dotoli, M. P. Fanti, and A. M. Mangini, "Real Time Identification of Discrete Event Systems by Petri Nets," in *IFAC 2007*, 2007.

[8] F. Basile, S. Member, P. Chiacchio, S. Member, and J. Coppola, "Identification of Time Petri Net Models," *IEEE Trans. Syst. Man Cybern. Syst.*, pp. 1–15, 2016.

[9] A. Rozinat, R. S. Mans, M. Song, and W. M. P. van der Aalst, "Discovering simulation models," *Inf. Syst.*, vol. 34, no. 3, pp. 305–327, 2009.

[10] M. P. Cabasino, "Fault diagnosis and identification of discrete event systems using Petri nets," University of Cagliari, 2008.

[11] S. Ould El Mehdi, R. Bekrar, N. Messai, E. Leclercq, D. Lefebvre, and B. Riera, "Design and Identification of Stochastic and Deterministic Stochastic Petri Nets," *IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans*, vol. 42, no. 4, pp. 931–946, 2012.

[12] R. Poli, W. B. Langdon, N. F. Mcphee, and J. R. Koza, "A Field Guide to Genetic Programming," 2008.

[13] D. P. Searson, "GPTIPS 2: an open-source software platform for symbolic data mining," 2015.

Denno, Peter; Kulkarni, Amogh; Balasubramanian, Daniel; Karsai, Gabor.
"Production System Identification with Genetic Programming."
Paper presented at 2018 IEEE International Conference on Industrial Technology (ICIT), Lyon, France. February 19, 2018 - February 22, 2018.

ASME's Power and Energy Conference and Exposition
PowerEnergy2018
June 24-28, 2018, Lake Buena Vista, Florida, USA

PowerEnergy2018-7295

ENERGYPLUS INTEGRATION INTO CO-SIMULATION ENVIRONMENT TO IMPROVE HOME ENERGY SAVING THROUGH CYBER-PHYSICAL SYSTEMS DEVELOPMENT

Joe Singer*, Thomas Roth, Chenli Wang*, Cuong Nguyen**, Hohyun Lee***

*Santa Clara University
Department of Mechanical Engineering
Santa Clara, California, USA, 95053

**National Institute of Standards and Technology,
Gaithersburg, MD, USA, 20899

ABSTRACT

This paper presents a co-simulation platform which combines a building simulation tool with a Cyber-Physical Systems (CPS) approach. Residential buildings have a great potential of energy reduction by controlling home equipment based on usage information. A CPS can eliminate unnecessary energy usage on a small, local scale by autonomously optimizing equipment activity, based on sensor measurements from the home. It can also allow peak shaving from the grid if a collection of homes are connected. However, lack of verification tools limits effective development of CPS products. The present work integrates EnergyPlus, which is a widely adopted building simulation tool, into an open-source development environment for CPS released by the National Institute of Standards and Technology (NIST). The NIST environment utilizes the IEEE High Level Architecture (HLA) standard for data exchange and logical timing control to integrate a suite of simulators into a common platform. A simple CPS model, which controls local HVAC temperature set-point based on environmental conditions, was tested with the developed co-simulation platform. The proposed platform can be expanded to integrate various simulation tools and various home simulations, thereby allowing for co-simulation of more intricate building energy systems.

INTRODUCTION

Heating, ventilation and cooling (HVAC) is the largest source of residential energy consumption in United States, encompassing about 25% of total residential energy usage. A significant portion of energy is wasted by unnecessary operation, such as overheating/overcooling or operation without occupants. Based on the recent paper by Nguyen and Aiello [1], energy conscious behaviors in residential homes can lead to 33% less energy consumption compared to the design point, and 50% less energy consumption compared to those demonstrating wasteful behaviors. Incorporating intelligence into home appliances to make them Cyber-Physical Systems (CPS) can allow wasteful devices to exhibit energy conscious behaviors. CPS are systems involving interactions between computation and physical components that process information collected from sensors to control physical actuators [2]. For example, a smart HVAC system can consume less energy by replacing fixed operating set-points with dynamic operation based on occupancy sensor information. Assuming residential homes achieve 33% consumed energy reduction through CPS appliance operation, an upwards of 3.6 quadrillion BTUs of energy can be saved nationally in the United States' residential sector [3].

Several products currently manipulate appliance operation via CPS paradigm [4, 5]. Unfortunately, their functionality is limited to remote controllability because implementation of automated control algorithms is challenging. Physical building interactions required complex testing and validation that cannot be captured in a single CPS component. Such experimental CPS necessitates interdisciplinary knowledge and real-time data collection, which requires significant amounts of time and resources [6]. A co-simulation platform involving a building simulator that combines multiple CPS components can address the aforementioned challenges in developing and implementing automated building controls.

The National Institute of Standards and Technology (NIST) developed an open-source CPS experiment and testing environment called Universal CPS Environment for Federation (UCEF). Its graphical user interface is designed to make co-simulations and experiments for CPS product simple and available. UCEF integrates various simulation entities (federates) sourced in different development environments, which has traditionally been challenging to accomplish [7]. UCEF leverages the IEEE's High Level Architecture (HLA) standard [8] for its communication protocol, implemented by the

Roth, Thomas; Nguyen, Cuong; Singer, Joe; Wang, Chenli; Lee, Hohyun.
"EnergyPlus Integration into Co-Simulation Environment to Improve Home Energy Saving Through Cyber-Physical Systems Development."
Paper presented at ASME Energy Sustainability Conference, Lake Buena Vista, FL, United States. June 24, 2018 - June 28, 2018.

Portico Run-Time Infrastructure (RTI) [9], to achieve logical timing control and data transfer among a collection of federates (defined as a federation). UCEF remains under development and supports few simulation tools. Integrating a building specific simulation tool into UCEF as a new supported federate type can lead to simpler building automation experimentation and validation.

An open-source building simulation tool called EnergyPlus [10] can be integrated with UCEF to achieve data exchange in an CPS test environment. EnergyPlus, a widely used modeling tool created by the Department of Energy, can evaluate building energy consumption at sub-hourly time steps. The software calculation capabilities incorporate key building parameters such as user activity, HVAC systems, building composition, and more. In CPS testing, EnergyPlus can replace physically measured building information with a simulated model to accelerate CPS tool development.

The present work integrates EnergyPlus into UCEF. An EnergyPlus model will communicate building information to the RTI using a UCEF Java federate. Although this work can benefit any building control system, a simple model was tested to verify co-simulation capability. A HVAC set-point algorithm implemented in another Java federate will receive environment temperature from EnergyPlus and return HVAC set-points to EnergyPlus. Intelligent set-point control of an HVAC system can significantly reduce energy consumption in a residential building between 33%-50% [1, 11], providing a good use case for the developed platform. Further, other simulators integrated with UCEF can expand HVAC controllability to include pre-heating or pre-cooling a collection of homes to reduce excessive power draw during peak demand [12]. Enabling UCEF co-simulation with EnergyPlus creates a simple approach to reduce residential energy consumption by allowing automation and optimization of wasteful appliances.

ACRONYMNS

BTU	British Thermal Unit
CPS	Cyber-Physical Systems
FMI	Functional Mockup Interface
FMU	Functional Mockup Unit
HLA	High Level Architecture
HVAC	Heating, Ventilation, and Air Conditioning
IDF	Input Data File
IEEE	Institute of Electrical and Electronic Engineers
NIST	National Institute of Standards and Technology
RTI	Run-Time Infrastructure
TCP/IP	Transmission Control Protocol/Internet Protocol
UCEF	Universal CPS Environment for Federation
XML	Extensible Markup Language

APPROACH

EnergyPlus currently has an existing co-simulation interface through the Functional Mock-up Interface (FMI) standard created by Modelisar [13]. The standard accomplishes interoperability by connecting simulation platforms to an external model by use of a zip file (with extension **.fmu*) known as a Functional Mock-up Unit (FMU). The zip file contains three elements: an Extensible Markup Language (XML) file, compiled C code binaries, and optional documentation for data exchange. The XML file establishes interfacing data, the C code orchestrates data exchange, and the documentation can define and specify operation. Unfortunately, the FMI and HLA standards have different notions of time management, and UCEF does not support data exchange using FMI. To bridge the two standards, we create an FMU with capabilities for bi-directional communication between an EnergyPlus model and a UCEF Java federate. The Java federate will be customized to wrap EnergyPlus models for data exchange to an RTI federation. This data communication, represented in Figure 1, is done through TCP/IP socket communication between our FMU and Java federate.

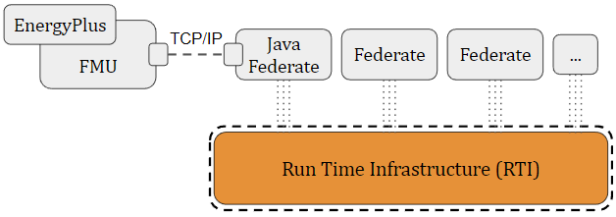


Figure 1: EnergyPlus has capability to interface with a FMU. Using TCP/IP socket communication inside a generic FMU allows for connectivity to a UCEF Java federate for the HLA RTI data exchange.

Connecting EnergyPlus to an FMU involves specific modifications of an EnergyPlus input data file (IDF). An IDF defines parameters to perform building energy simulations, such as building materials, components, and equipment. Using an IDF component called *FunctionalMockupUnitImport*, co-simulation is linked between EnergyPlus and the FMU. This component initializes the FMI master and slave architecture where slaves are coordinated and executed by the master program. EnergyPlus acts as the master in this configuration, which initializes the FMU as an executable slave instance.

Upon simulation start with the aforementioned component, EnergyPlus locates and unpacks the linked FMU to begin processes represented in Figure 2. Execution of the FMU's customized C binaries is controlled by EnergyPlus to run select FMI functions [14] that have been modified and implemented to exchange data with the HLA RTI. EnergyPlus first calls the *fmiInstantiateSlave* function to parse through the unpacked XML file to properly allocate memory for interface data. Next, the *fmiInitializeSlave* function uses TCP/IP sockets to establish connection to a server hosted in the UCEF Java federate. After TCP/IP connection is verified, EnergyPlus's time step calculations begin.

Roth, Thomas; Nguyen, Cuong; Singer, Joe; Wang, Chenli; Lee, Hohyun.
"EnergyPlus Integration into Co-Simulation Environment to Improve Home Energy Saving Through Cyber-Physical Systems Development."
Paper presented at ASME Energy Sustainability Conference, Lake Buena Vista, FL, United States. June 24, 2018 - June 28, 2018.

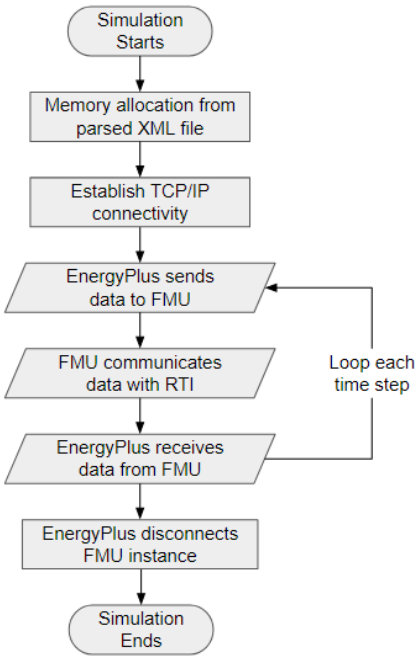


Figure 2: EnergyPlus as a master program for the Functional Mockup Interface (FMI) standard calls select functions throughout simulation to perform specific tasks. At each time step, three functions are called to transfer EnergyPlus data to a Functional Mockup Unit (FMU).

At each time step, EnergyPlus sends data to the FMU as a *real* data type using the function *fmiSetReal*. The FMU will then utilize socket connection in *fmiDoStep* to send the EnergyPlus data (as a concatenated string) to the designated Java federate. The format of this string is standardized and represented as follows:

HEADER\r\nTIMESTAMP\r\nNAME\r\nVALUE\r\n….
NAME\r\nVALUE\r\n\r\n

The "HEADER" defines handling procedures of the string. Data sent from FMU to the Java federate will either contain the header "UPDATE" or "TERMINATE". An "UPDATE" header is used at each EnergyPlus time step to signify incoming information to the Java federate. A "TERMINATE" header informs the Java federate that EnergyPlus simulation has ended. Data received by the FMU from the Java federate will either contain "SET" or "NOUPDATE" headers. "SET" indicates federation interactions will change EnergyPlus variables, and "NOUPDATE" indicates no variables will change. After the header, the "TIMESTAMP" communicates simulation time (in seconds) for logical time management. Next, for each "UPDATE" and "SET" header, "NAME" and "VALUE" respectively represent each variable name and corresponding value of interfacing data defined through the XML file. Each piece of information is separated by a carriage return followed by a line feed ("\r\n"). Two consecutive cartridge returns and line feeds at the end signify the end of the string.

EnergyPlus will remain in the *fmiDoStep* function until the Java federate responds with a concatenated string. After a string is returned, the FMU will parse through the returned string in *fmiDoStep*. The *fmiGetReal* function passes received information back into the EnergyPlus model as a *real* data type. The described data exchange pipeline is represented in Figure 3. The master EnergyPlus program will exchange data with the Java federate at each time step. After the final time step, the FMU slave instance is disconnected, and the simulation ends.

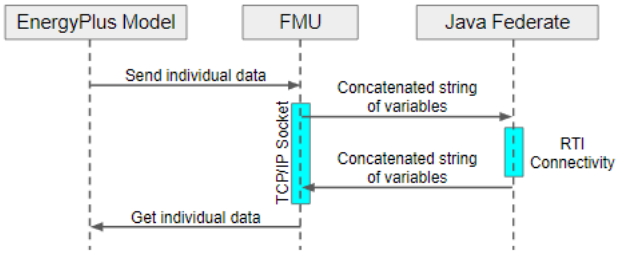


Figure 3: UML diagram representing data communication between the master EnergyPlus program and a UCEF Java federate via FMU slave instance.

The Java federate developed in UCEF communicates information between the FMU slave instance and the RTI. This federate begins by hosting a TCP/IP server for the FMU client connection. During simulation, the federate parses each received string from the FMU and passes its information to the RTI federation. The federate then waits for messages from the RTI that should be sent to EnergyPlus. A concatenated string containing the content of these messages, is then returned to the FMU client.

CONNECTIVITY VALIDATION

A series of simulations were executed to validate EnergyPlus communication with an HLA RTI federation. A simple three-room house model was created in an EnergyPlus IDF. The home was modeled as being located in San Francisco, CA, USA using weather information from June 2017. This location is chosen because hourly temperature change is high enough to potentially cause unnecessary heating or cooling. The home was equipped with a dual set-point HVAC system operating at a temperature range between 21□ and 23□ to simulate arbitrary occupant comfort range. The first simulation executed the standard EnergyPlus model without the implemented FMU external interface. Environmental temperature, zone temperature, and HVAC energy usage information were recorded at each time step. Resulting HVAC energy consumption using these "naive" set-points can resemble non-energy conscious behaviors.

The second simulation directly supplied environment temperature data through to a standalone thermostat controller algorithm. A simple algorithm in Java was created to adjust

Roth, Thomas; Nguyen, Cuong; Singer, Joe; Wang, Chenli; Lee, Hohyun.
"EnergyPlus Integration into Co-Simulation Environment to Improve Home Energy Saving Through Cyber-Physical Systems Development."
Paper presented at ASME Energy Sustainability Conference, Lake Buena Vista, FL, United States. June 24, 2018 - June 28, 2018.

heating and cooling temperature set-points based on environment temperature. When outside temperature is either high or low, the heating and cooling setpoints can adjust correspondingly to minimize HVAC loading and to preheat/precool the house. EnergyPlus and UCEF were not used in this second simulation. Rather, environment temperature recorded in the first simulation was directed through the thermostat controller algorithm to return dynamic dual set-points. The results of this second experiment act as a ground truth for the next experiment implementing EnergyPlus and UCEF connectivity.

The final simulation implemented the FMI external interface with the IDF described in the first simulation. The *FunctionalMockupUnitImport* class enforced data exchange with our developed FMU, linking EnergyPlus to our Java federate. The federation was created using UCEF, binding EnergyPlus and the secondary thermostat controller algorithm through RTI. Shown in Figure 4, environment temperature from EnergyPlus was sent through RTI to the thermostat controller at each time step. Before time step progression, the controller returned a heating and cooling set-point to the HVAC system in EnergyPlus. HVAC set-points and zone temperature were recorded.

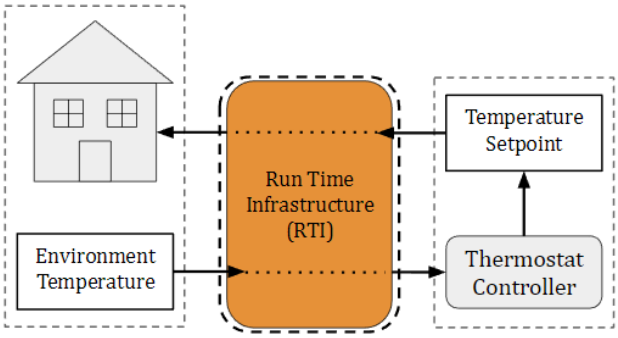


Figure 4: Representation of the data transfer using Run Time Infrastructure between the EnergyPlus Java Federate and the thermostat controller Java federate.

RESULTS & DISCUSSION

The first simulation recorded sub-hourly temperature and HVAC energy consumption data of a simple EnergyPlus model. Dual set-point of an HVAC system between 21□ and 23□ caused activation. Heating activated in the morning and evening, and cooling activated mid-day. Figure 5 shows the zone temperature fluctuation throughout the design day. Figure 6 represents respective heating and cooling energy consumed by the system. HVAC operation between the narrow temperature range represents excess consumed energy by a non-energy conscious user. The following simulations attempt to incorporate in intelligent CPS to control the model HVAC.

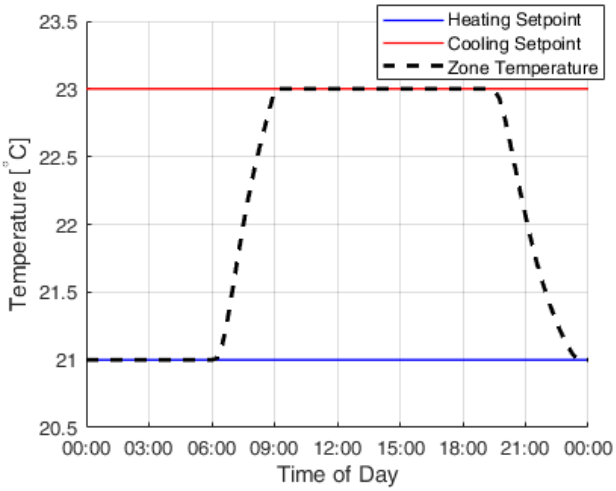


Figure 5: Simulation results of temperature fluctuation controlled by constant heating and cooling setpoints input.

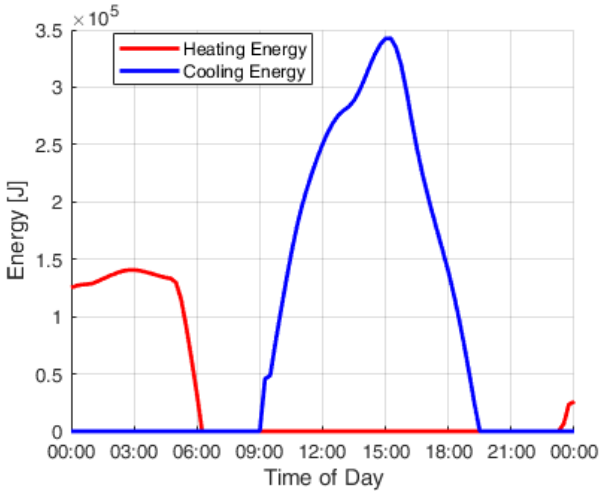


Figure 6: Heating and cooling power consumptions respective to HVAC operation controlled by setpoints (Figure 5).

A thermostat controller output dynamic set-points based on environment temperature. Direct input of the second simulation and EnergyPlus/RTI input of the third simulation yielded identical results, shown in Figure 7. Matching outputs of the two simulations validates continuous and accurate EnergyPlus integration with UCEF.

Roth, Thomas; Nguyen, Cuong; Singer, Joe; Wang, Chenli; Lee, Hohyun.
"EnergyPlus Integration into Co-Simulation Environment to Improve Home Energy Saving Through Cyber-Physical Systems Development."
Paper presented at ASME Energy Sustainability Conference, Lake Buena Vista, FL, United States. June 24, 2018 - June 28, 2018.

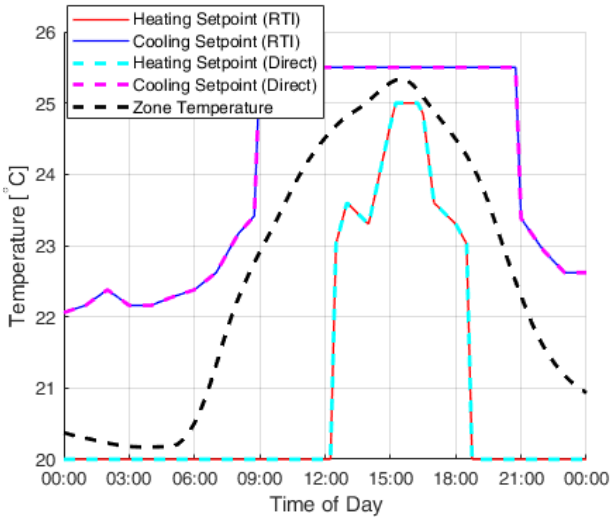


Figure 7: HVAC heating and cooling set-points based on external thermostat controller. Direct connection based and RTI connection yield consistent results.

Figure 7 also shows internal zone temperature of the EnergyPlus model. Dynamic thermostat controller outputs cause zero energy consumption for this simulation day. Compared to the naive set-points of the first simulation occupant, EnergyPlus co-simulation with the intelligent thermostat controller removed unnecessary energy consumption. Results verify UCEF integration does not impact simulated results.

CONCLUSION & FUTURE WORK

This paper developed an open-source integration of a building simulation software with UCEF for the design and validation of CPS. By developing a simple FMU with a TCP/IP connection to a Java federate, calculated data at each time step was communicated from an EnergyPlus model to an HLA federation. This successful integration allows co-simulation between EnergyPlus models and HLA federates. Simulated results validate UCEF-based federations can exchange data with EnergyPlus models without negative impact on results. More complex control algorithms and other simulation tools integrated into EnergyPlus creates an environment that can produce sophisticated CPS that reduce energy consumption in residential buildings. Integration of EnergyPlus into UCEF as a new federate type enhances the platform's capabilities through added support of building simulations.

Additional concepts can be further investigated for more robust development. Modifications of FMU configuration files may be necessary for different simulation designs requiring different building model information. Currently, the IDF and the XML file need to be created manually based on the desired interface data. UCEF has support for the automatic generation of configuration files based on the content of fields in its graphical user interface. A user should be able to enter desired EnergyPlus variable information directly into the UCEF interface to automatically generate and update the IDF and XML file, rather than having to write the files themselves. Future work could address this usability feature through extensions to the UCEF graphical interface.

The presented approach using TCP/IP sockets could be further leveraged to integrate other FMI tools into UCEF. FMUs connected to other programs can utilize the TCP/IP concatenated string protocol to communicate with the Java federate in UCEF. Expanding co-simulation diversity to FMI connected tools can vastly improve UCEF simulator and emulator inventory. UCEF integration can increase development effectivity by allowing for improved logical timing control of these FMI tools.

ACKNOWLEDGMENTS

Portions of this publication and research efforts are made possible through the support of NIST via federal award #70NANB17H039.

Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain commercial products are identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose.

REFERENCES

[1] Nguyen, T. A., and Aiello, M., 2013, "Energy intelligent buildings based on user activity: A survey," Energy and Buildings, 56, pp. 244-257.

[2] Griffor, E. R., Greer, C., Wollman, D. A., and Burns, M. J., June 2017, "Framework for Cyber-Physical Systems: Volume 1, Overview," https://dx.doi.org/10.6028/NIST.SP.1500-201.

[3] Lawrence Livermore National Laboratory, 2016, "Energy Flow Charts: Charting the Complex Relationships among Energy, Water, and Carbon," https://flowcharts.llnl.gov/.

[4] Alhafidh, B. M. H., and Allen, W. H., 2017, "High Level Design of a Home Autonomous System Based on Cyber Physical System Modeling," IEEE, p. 45.

[5] Hong, T., Sun, H., Chen, Y., Taylor-Lange, S. C., and Yan, D., 2016, "An occupant behavior modeling tool for co-simulation," Energy & Buildings, 117, pp. 272-281.

[6] Terpening, E. D., and Littleton, A., 2017, "The State of Internet of Things in the Home: PART II: OPPORTUNITIES AND CHALLENGES FOR BRANDS SELLING IOT PRODUCTS FOR THE HOME," Altimeter Group - Research Reports, p. 1.

[7] Roth , T., Song , E., Burns , M., Neema , H., Emfinger, W., and Sztipanovits, J., 2017, "Cyber-Physical System Development Environment for Energy Applications," ASME

Roth, Thomas; Nguyen, Cuong; Singer, Joe; Wang, Chenli; Lee, Hohyun.
"EnergyPlus Integration into Co-Simulation Environment to Improve Home Energy Saving Through Cyber-Physical Systems Development."
Paper presented at ASME Energy Sustainability Conference, Lake Buena Vista, FL, United States. June 24, 2018 - June 28, 2018.

2017 11th International Conference on Energy Sustainability, Charlotte, North Carolina, USA.
[8] IEEE, 2010, "IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)-- Framework and Rules," IEEE Std 1516-2010, pp. 1-38.
[9] Pokorny, T., and Fraser, M., 2017, "The Portico Project," http://timpokorny.github.io/public/index.html.
[10] US Department of Energy's Building Technologies Office, and National Renewable Energy Laboratory (NREL), 2017, "EnergyPlus," https://energyplus.net/.
[11] Hong, T., and Lin, H.-W., 2013, "Occupant Behavior: Impact on Energy Use of Private Offices," Conference: ASim 2012 - 1st Asia conference of International Building Performance Simulation Association, Shanghai, China.
[12] Liu, Y., Qiu, B., Fan, X., Zhu, H., and Han, B., 2016, "Review of Smart Home Energy Management Systems," Energy Procedia, 104, pp. 504-508.
[13] Junghanns, A., 2017, "FMI Functional Mock-up Interface," http://fmi-standard.org/.
[14] Modelisar, 2010, "Functional Mock-up Interface for Co-Simulation v1.0 (Documentation)."

Roth, Thomas; Nguyen, Cuong; Singer, Joe; Wang, Chenli; Lee, Hohyun.
"EnergyPlus Integration into Co-Simulation Environment to Improve Home Energy Saving Through Cyber-Physical Systems Development."
Paper presented at ASME Energy Sustainability Conference, Lake Buena Vista, FL, United States. June 24, 2018 - June 28, 2018.

Assessment of Radiation Solvers for Fire Simulation Models Using RADNNET-ZM

Wai Cheong Tam[a,*], Walter W. Yuen[b]

[a]*Fire Research Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA*
[b]*Department of Mechanical Engineering, University of California at Santa Barbara, Santa Barbara, California, USA*

Abstract

The paper presents a neural-network based zonal method (RADNNET-ZM) for the analysis of radiative heat transfer in an arbitrary Cartesian enclosure with an isothermal, homogeneous, non-gray medium. The model accounts for the non-gray effect of absorbing species in a combustion environment and the geometric effect of any three-dimensional enclosures. The model is verified against benchmark solutions. Maximum local error is observed to be less than 4 %. Prediction accuracy of an existing zonal radiation solver is assessed. Results demonstrate that RADNNET-ZM can provide substantial improvement to zone fire simulation models for the prediction of radiative heat transfer without a significant increase in computation cost.

Keywords: Neural network, zonal method, non-gray, multi-dimensional, fire simulation model

Nomenclature

a_λ	local absorption coefficient
A_i	elemental area *i*
D	grid size of discretization
$e_{\lambda b}$	Planck function
f_v	soot volume fraction
F_{ij}	view factor between A_i and A_j
$F_{ss,xx}$	generic exchange factor (xx = pd, pp)
$L_{ij,xx}$	center-to-center distance between A_i and A_j (xx = pd, pp)
$L_{pd,x}$	mean beam length between two perpendicular elemental areas (x = soot, gas)
L_{pp}	mean beam length between two parallel elemental areas
n_x,n_y,n_z	dimensionless distances for A_j relative to A_i
P_{CO_2}	partial pressure of CO_2
P_{H_2O}	partial pressure of H_2O
P_g	total pressure of an $N_2/H_2O/CO_2$ mixture
$\dot{q}"_g$	incident heat flux due to emission of mixture medium
$\dot{q}"_w$	incident heat flux due to emission of wall
ss	surface-surface exchange factor
SS	total surface-surface exchange factor
T_g	gas temperature
T_w	wall temperature
X_{CO_2}	mole fraction of CO_2
X, Y, Z	dimensions of an enclosure
Greek symbols	
α	total absorptivity (sum of soot and gas absorptivity)
α_s	soot absorptivity
$\Delta\alpha$	gas absorptivity
β_{xx}	normalized mean beam length (xx = pd, pp)
λ	wavelength
ε	emissivity of gas mixture
σ	Stefan-Boltzmann constant
Subscripts	
pd	perpendicular
pp	parallel

* Corresponding author. Tel.: +1- 301-975-8202.
E-mail address: waicheong.tam@nist.gov

Tam, Wai Cheong; Yuen, Walter.
"Assessment of Radiation Solvers of Fire Simulation Models Using RADNNET-ZM."
Paper presented at 11th Asia-Oceania Symposium on Fire Science and Technology (AOSFST), Taipei, Taiwan Province of China. October 21, 2018 - October 25, 2018.

1. Introduction

A significant amount of research has been conducted in both computational techniques for multi-dimensional radiation heat transfer and the understanding of spectroscopic absorption properties of different combustion gases over the past 30 years. To account for the geometric effect, there are zonal methods [1], discrete ordinate methods [2], discrete transfer methods [3], and many others [4]. To simulate the spectral effect, there are narrow band models [5], k-distribution models [6], and weighted sum of gray gas models [7]. In recent years, simulation methods to simultaneously account for both the spectral and geometric effect have also been developed [8]. Despite these efforts to deal with the non-gray multi-dimensional aspects of radiative heat transfer, few have been utilized in any significant degree by the engineering design and fire safety community in the areas of combustion and fire, where in many cases the effect of radiation is known to be not only important, but dominant. The primary difficulty is the mathematical complexity.

The evaluation of radiative heat transfer with the presence of a participating medium consisting of typical combustion products (H_2O, CO_2, and soot particulate) at low pressure (i.e., one atmosphere) in a three-dimensional enclosure is numerically complex. For a one-dimensional isothermal homogeneous medium, the absorptivity is a complicated function of six independent variables (optical thickness of H_2O, CO_2, and CO, source temperature, mixture temperature, and soot volume fraction) [9]. To obtain an accurate evaluation of the spectral behavior of the gas mixture, a direct numerical integration using realistic spectral data (i.e., the narrow band model) is required to be carried out. For the evaluation of radiative heat transfer between surfaces and the gaseous medium, the geometric effect is significant and another direct numerical integration is required. In a typical calculation, such as to simulate the transient thermal environment within a fire resistance furnace, previous work [10] indicates that more than 60 million numerical evaluations are needed to determine the exchange factors. Since the condition within the furnace is continuously changing, the exchange factors are required to be re-evaluated for every time step in a simulation. This level of computational effort is clearly not feasible for practical engineering applications. For this reason, many existing zone fire models, such as CFAST (Consolidated Fire And Smoke Transport [11]), and CFD codes, including FDS (Fire Dynamics Simulator [12]) and FLUENT [13], implement approximate radiation solvers which rely on empirical charts/correlations to enhance computational efficiency.

The objective of this paper is to present a generalized radiation solver that has the capabilities to simulate accurately the realistic effect of radiation heat transfer in any arbitrary three-dimensional fire/combustion environment efficiently. The radiation solver, RADNNET-ZM (RADiation Neural NETwork-Zonal Method), is a generalization of the zonal method using the concept of the generic exchange factor (GEF) [14]. The GEFs have been demonstrated to be an efficient approach for the evaluation of radiative heat transfer in a multi-dimensional gray medium. In a recent work [15], the GEFs are expanded to account for the non-gray effect of a H_2O/CO_2/soot mixture medium by using a neural network correlation, RADNNET [9]. To improve the computational efficiency in the evaluation of the GEFs, the concept of mean beam length (MBL) is utilized. Additional neural networks are generated for the MBLs to characterize the non-gray absorptivity and emissivity of the medium bounded by the elemental surfaces. Using superposition, total exchange factors between arbitrary surfaces can be determined for the evaluation of radiative heat transfer within an enclosure accurately and efficiently.

In the following sections, the mathematical formulation of RADNNET-ZM is presented. Model verification for RADNNET-ZM is provided. Comparison with results generated from the radiation solver used in CFAST [11] is made. Note that the development of the GEFs is described with details in [14, 15]. Therefore, only the features related to the incorporation of the GEFs in the solution algorithms are addressed below.

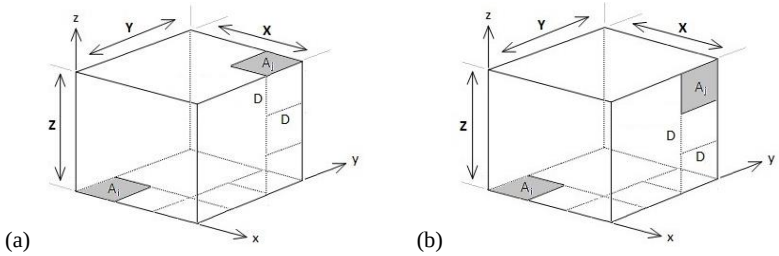


Fig 1. Geometry for (a) parallel surface-surface GEF, $F_{ss,pp}$ and (b) perpendicular surface-surface GEF, $F_{ss,pd}$.

2. Mathematical formulation

Consider a one-zone enclosure filled with a mixture of water vapor, carbon dioxide, and soot particulate with arbitrary dimensions of X, Y, and Z as shown in Fig. 1, the analysis of radiative heat transfer to the bounding surfaces requires the evaluation of surface–surface exchange factors. Mathematically, the surface–surface exchange factors between two surfaces, $s_i s_j$, can be evaluated from the following integration [8]

Tam, Wai Cheong; Yuen, Walter.
"Assessment of Radiation Solvers of Fire Simulation Models Using RADNNET-ZM."
Paper presented at 11th Asia-Oceania Symposium on Fire Science and Technology (AOSFST), Taipei, Taiwan Province of China. October 21, 2018 - October 25, 2018.

$$s_i s_j = \frac{1}{\sigma T_w^4} \int_{A_i} \int_{A_j} \int_0^\infty \frac{e_{\lambda b}(T_w) e^{-a_\lambda (T_g) S} \cos\theta_i \cos\theta_j}{\pi S^2} d\lambda dA_j dA_i \quad (1)$$

where T_w is the surface temperature of the source area A_i, T_g is the gas temperature, S is the line-of-sight distance between the two integration area elements, dA_i and dA_j, θ_i and θ_j are the angles between the line-of-sight and the unit normal vector of the two differential area elements, σ is the Stefan-Boltzmann constant, $e_{\lambda b}$ is the blackbody emissive power, and a_λ is the local absorption coefficient. To achieve good accuracy, approximately 60 million numerical evaluations of Eq. (1) are required for one time-step in a typical fire simulation.

Using the concept of GEF, the expression for the evaluation of the surface–surface exchange factors is reduced into a simple form and the exchange factors can be determined from existing neural network correlations such that the numerical evaluations of Eq. (1) will no longer have to be performed during the actual calculation. Specifically, the surface-surface exchange factor for two parallel elemental areas, $s_i s_{j,pp}$, as shown in Fig. 1a is

$$s_i s_{j,pp} = D^2 F_{ss,pp}\left(T_w, T_g, P_g D, X_{CO_2}, f_v D, n_x, n_y, n_z\right) \quad (2)$$

where D is the grid size for the discretization and $F_{ss,pp}$ is the GEF. For simplicity, the mathematical formulation is described under the assumption that the boundary of the enclosure can be subdivided into square elements. Additional treatment is required to handle conditions with imperfect discretization and they will be presented in future publications. ($n_x = S_x/D$, $n_y = S_y/D$, $n_z = S_z/D$) are the non-dimensionalized locations of the receiving area A_j relative to the source area A_i. S_x, S_y, and S_z are the scalar components of the vector connecting the center points of the area elements. As an example, (n_x, n_y, n_z) is (3, 3, 3) for the two parallel elemental areas being shown in Fig. 1a. Based on Eq. (2), $F_{ss,pp}$ is a function of geometry and five combustion parameters: surface temperature (T_w), gas temperature (T_g), optical thickness (P_gD), mole fraction of CO_2 (X_{CO_2}), and soot volume fraction (f_v). The total pressure of the gas mixture is given as

$$P_g = P_{H_2O} + P_{CO_2} \quad (3)$$

and the mole fraction of CO_2 is

$$X_{CO_2} = \frac{P_{CO_2}}{P_{H_2O} + P_{CO_2}} \quad (4)$$

Using the concept of mean beam length (MBL), the GEF can be written in a one-dimensional form

$$F_{ss,pp} = F_{ij}\left(n_x, n_y, n_z\right)\left[1 - \alpha\left(T_w, T_g, P_g L_{pp}, X_{CO_2}, f_v L_{pp}\right)\right] \quad (5)$$

where F_{ij} is the view factor between the two elemental areas, α is the total absorptivity of the sooty gas mixture, and L_{pp} is the MBL accounting for mixture absorption. The numerical evaluation for the view factor is costly when the two areas are close to each other (i.e. $n_x \le 5$, $n_y \le 5$, $n_z \le 5$). For numerical efficiency, view factors are tabulated for ($n_x \le 5$, $n_y \le 5$, $n_z \le 5$) and the tabulated view factors are used to obtain the generic exchange factor. When n_x, n_y, or n_z is larger than 5, the center-to-center distance between the two elemental areas is used as the MBL for the evaluation of the view factors analytically and numerical experiments show that the error associated to the approximate view factor (using the center-to-center distance) is less than 1 %. α is obtained from RAD-NNET, a neural network correlation that predicts the one-dimensional total absorptivity for a $N_2/H_2O/CO_2$/soot mixture [9]. Given a combustion environment which can be described by the five combustion parameters, the neural network correlation provides the corresponding total absorptivity through "look-up tables". The absorptivity data are generated using RADCAL [4] for 550 discrete values of P_gD, 11 discrete values of X_{CO_2}, 10 discrete value of f_vD, and 18 discrete values of T_w and T_g, respectively, corresponding to a set of over 19 million data points. The ranges of input variables are

$$\begin{aligned} &0 \le P_g D \le 1000 \text{ kPa-m} \\ &0 \le X_{CO_2} \le 1 \\ &0 \le f_v D \le 10^{-6} \text{ m} \\ &300 \le T_w,\ T_g \le 2000 \text{ K} \end{aligned} \quad (6)$$

As shown in [9], the relative error associated with RAD-NNET is less than 5 % for absorptivity or emissivity values larger

than 0.01. Thus, RAD-NNET is expected to have the same order of accuracy as compared to RADCAL. It should be noted that the RAD-NNET prediction capability for the total absorptivity can continuously be improved with the inclusion of more numerical data. For the determination of the MBLs, Yuen et al. [15] demonstrate that the effect of combustion parameters is minor. For ($n_x = 1$, $n_y = 1$, $n_z = 1$) where the effect of combustion parameters to the MBL is observed to be the largest, the absolute values of the MBL vary about 9 % as a function of the combustion parameters as shown in Eq. (6). An average MBL is shown to be sufficient to yield an accurate value for the GEF in a specific dimensionless distance. For that, L_{pp} is given as

$$L_{pp} = L_{ij,pp}\beta_{pp} \tag{7}$$

where β_{pp} is the average normalized MBL for (n_x, n_y, n_z) and $L_{ij,pp}$ is the center-to-center distance between the two parallel elemental areas. As an example, β_{pp} for two parallel areas D^2 separated by a distance D, which is being denoted as ($n_x = 1$, $n_y = 1$, $n_z = 1$), is determined to be 1.10.

The surface-surface exchange factor for two perpendicular elemental areas, $s_i s_{j,pd}$, as shown in Fig. 1b can be expressed in a similar form of Eq. (1)

$$s_i s_{j,pd} = D^2 F_{ss,pd}\left(T_w, T_g, P_g D, X_{\mathrm{CO}_2}, f_v D, n_x, n_y, n_z\right) \tag{8}$$

with

$$F_{ss,pd} = F_{12}\left(n_x, n_y, n_z\right)\left[1-\alpha_s\left(T_w, f_v L_{pd,s}\right)-\Delta\alpha\left(T_w, T_g, P_g L_{pd,g}, X_{\mathrm{CO}_2}, f_v L_{pd,g}\right)\right] \tag{9}$$

where $F_{ss,pd}$ is the GEF for two perpendicular elemental areas and the total absorptivity, α, is separated into a soot component, α_s, and a gas component, $\Delta\alpha$. It is observed from numerical experiments conducted in [15] that the mathematical behavior of the MBL for perpendicular areas, L_{pd}, is more complex and L_{pd} varies strongly with the combustion parameters. For ($n_x = 1$, $n_y = 1$, $n_z = 1$), the MBL varies more than 65 % over the entire range of combustion parameters. For that, two MBLs, a soot MBL ($L_{pd,s}$) and a gas MBL ($L_{pd,g}$), are needed to account for the corresponding absorption characteristics accurately. $L_{pd,s}$ is tabulated for ($n_x \leq 4$, $n_y \leq 4$, $n_z \leq 4$) as a function of the product of surface temperature and soot concentration. If n_x, n_y, or n_z is out of range, the center-to-center distance between the two perpendicular elemental areas can be used. For $L_{pd,g}$, a neural network, MBLG-NNET (Mean Beam Length Gas – Neural NETwork), is generated for the ranges of the five combustion parameters. Geometrically, both MBLs are proportional to the center-to-center distance between the two areas and they are given by

$$L_{pd} = L_{ij,pd}\beta_{pd} \tag{10}$$

where β_{pd} is the corresponding normalized MBL for soot/gas mixture and $L_{ij,pd}$ is the center-to-center distance between the two perpendicular elemental areas.

In summary, with the use of the neural network correlations (RAD-NNET and MBLG-NNET), the numerical evaluations for solving Eq. (1) are not needed. Furthermore, it can be demonstrated that the exchange factors for any parallel/perpendicular elemental areas can be efficiently determined for any finite surfaces at any arbitrary locations in an enclosure with an isothermal, homogeneous, non-gray medium.

Utilizing the method of superposition, the total exchange factors between finite surfaces can be determined by summing over the generic exchange factors for all elemental areas. For any finite surface A_i in an enclosure, the incident radiative heat flux consisting of emission from the surrounding walls, $\dot{q}''_{w,i}$, and emission from the sooty gas mixture, $\dot{q}''_{g,i}$, are given by

$$\dot{q}''_{w,i} = \sigma T_{w,j}^4\left(1-\sum_{j=1}^{N} F_{ji}\alpha_{ij}\left(T_{w,j}, T_g, P_g R_x, X_{CO_2}, f_v R_x\right)\right) \tag{11}$$

$$\dot{q}''_{g,i} = \sigma T_g^4 \sum_{j=1}^{N} F_{ji}\alpha_{ij}\left(T_g, T_g, P_g R_x, X_{CO_2}, f_v R_x\right) \tag{12}$$

where N is the total number of bounding surfaces and R_x is the corresponding MBL for the different surfaces. Note that the details of the summation procedure and indexing associated with the superposition to the generic exchange factors in obtaining the total exchange factors in between two surfaces at arbitrary locations within an enclosure are demonstrated in [15]. For that, the details will not be provided in this paper. Readers can refer to the literature for detailed descriptions.

Tam, Wai Cheong; Yuen, Walter.
"Assessment of Radiation Solvers of Fire Simulation Models Using RADNNET-ZM."
Paper presented at 11th Asia-Oceania Symposium on Fire Science and Technology (AOSFST), Taipei, Taiwan Province of China. October 21, 2018 - October 25, 2018.

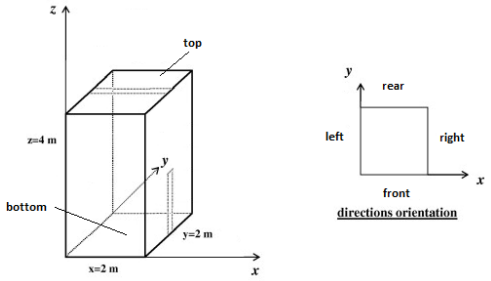


Fig. 2. Schematic of the considered geometry.

3. Results and discussion

3.1. Verification case: comparison between RADNNET-ZM and benchmark results

The predictive accuracy of the model described above is investigated. RADNNET-ZM is applied to a radiative heat transfer problem in a three-dimensional enclosure containing non-gray gases. Solutions for this problem were generated by Liu [16] using a statistical narrow band model (SNB) for the determination of the gas radiative properties and a ray tracing method [2] to solve for the radiative transfer equation (RTE). This benchmark problem has been used to verify newly developed/modified spectral models and/or solution methods [17]. Due to its reliable accuracy, the benchmark results will be used as the exact solutions for the verification process.

Fig. 2 shows the geometry being considered in this verification study. It is a three-dimensional rectangular enclosure with the dimensions of 2 m by 2 m by 4 m. The surrounding walls are assumed to be black. Surface temperature for all walls is maintained at 300 K. The total pressure of the gas mixture is kept at 1 atm. The temperature for the gas mixture is assumed to be isothermal and is maintained at 1000 K. The composition of the gas mixture is assumed to be homogeneous and contains pure water vapor. The verification conditions are summarized in Table 1.

The verification case is calculated using uniform grids of 11 by 11 by 11 and 16 by 16 by 16. Numerical results from RADNNET-ZM are generated. The predicted incident heat flux to various locations in the x-direction at the center of the y-direction on the top surface, denoted as (x, 1 m, 4 m), and the predicted incident heat flux at different locations in the z-direction at the center of the y-direction on the right surface, denoted as (2 m, 1 m, z), are obtained. Comparing the benchmark results generated by Liu, relative errors associated with RADNNET-ZM are determined. Due to symmetry condition of this problem, only half of the incident heat flux to the right surface will be presented. As shown in Fig. 3, it can be observed that the predicted results are in very good agreement with the benchmark results. The maximum local error is approximately 4 % and the large errors appeared at two ends (i.e. x = 0 m and x = 2 m as shown in Fig. 3a) are probably due to the angular discretization being used in Liu's study. For that, RADNNET-ZM solutions are possibly more accurate near the end of the wall. Overall, the results demonstrate that the RADNNET-ZM is capable to simulate both the spectral and geometric effect accurately.

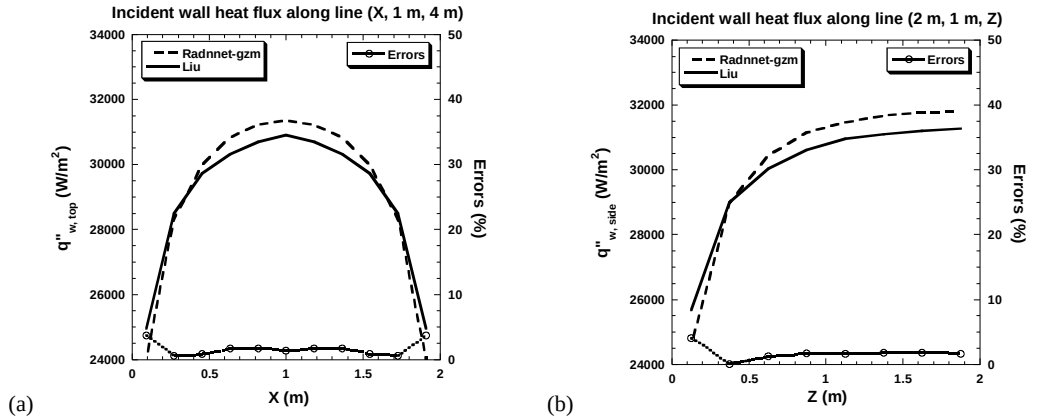


Fig. 3. Incident wall heat flux alone lines (a) [x, 1m, 4m] and (b) [2m, 1m, z] obtained from RADNNET-ZM and Liu [13] together with the error associated with RADNNET-ZM.

3.2. Assessment of CFAST radiation solver using RADNNET-ZM

CFAST [14] is a fire simulation model that divides compartments into two zones. Each zone includes a gas mixture/soot medium bounded by a ceiling or a floor, and four surfaces. Thermal conditions of each zone are assumed to be uniform. When there is a fire, a hot layer will form and the medium can be divided into an upper layer and a lower layer. If the fire persists, the upper layer, consisting of combustion by-products such as H_2O, CO_2, and soot particulate, increases in depth and the upper layer temperature increases. At high temperature, thermal radiation becomes the dominant mode of heat transfer between surfaces and the medium. For that, accurate evaluation of the radiative heat transfer for a participating medium in a three-dimensional enclosure becomes crucial. However, since the absorption behavior of real gases, such as H_2O and CO_2, is a strong function of wavelength, temperature, and species concentrations, brute force numerical evaluation to account for the effect of absorption and geometry is required as discussed in Ref. [10]. Yet, this approach is not feasible for practical engineering calculations. For this reason, in many zone fire models, including CFAST, simplifications are made to their radiation solvers for the evaluation of radiative heat transfer. To account for the non-gray spectral effect of real gases, a 1-D empirical correlation known as the Hottel's emissivity chart [1] is typically used. To account for the geometric effect, a constant MBL, based on some ad-hoc length scales without mathematical validation, is used. The application of these simplifications for the evaluation of radiative heat transfer in multi-dimensional non-gray media has not been validated. Therefore, the prediction accuracy associated with these radiation solvers is uncertain.

In the following section, the prediction accuracy of the radiation solver from CFAST is assessed. In the discussion given below, the CFAST radiation solver will be denoted as the "approximate approach". It should be noted that with the implementation of the generic exchange factor, bounding surfaces can be subdivided even in a one-zone calculation. It can also be demonstrated that RADNNET-ZM is able to capture local radiative heat transfer effect in a 3-D enclosure with an isothermal, homogeneous, non-gray medium accurately.

3.2.1. Test case 1: localized incident heat flux to surfaces (RADNNET-ZM vs. radiation solver using a constant MBL)

This case, identical to the verification case as shown in section 3.1, is simulated using the approximate approach. Target devices are specified on the top surface along (x, 1 m, 4 m) and the surface on the right along (2 m, 1 m, z) such that the corresponding incident heat flux can be obtained. Prediction generated from the approximate approach is compared to that obtained from RADNNET-ZM. Comparing RADNNET-ZM, relative errors associated with the approximate approach are also determined. As shown in Fig. 4, the approximate approach over-predicts the incident heat flux to the surface by as much as 28 %. This error is caused by the simplification used in the determination of mean beam length.

3.2.2. Test case 2: radiative properties of gas mixture (RADNNET-ZM vs. radiation solver using 1-D emissivity expression)

For black surfaces, the emissivity of the gas mixture, radiating to surface A_i, can be obtained from the solution to the analysis of an emitting medium and non-emitting wall [10] as

$$\varepsilon_g\left(T_g,T_g\right)=\frac{Q_i}{A_i\sigma T_g^4}\quad;\quad Q_i=\left[A_i-\sum_{j\neq i}S_iS_j\left(T_g,T_g\right)\right]\sigma T_g^4 \tag{13}$$

where Q_i is the radiative heat flux to surface A_i due to emission from the hot medium. S_iS_j is the total exchange factor and it can be determined by summing over the generic exchange factors for all elemental areas. With identical wall temperature for all surfaces, the mixture absorptivity due to wall emission is given to be

$$\alpha_g\left(T_w,T_g\right)=\frac{1}{A_i}\sum_{j\neq i}\left[A_iF_{ij}-S_iS_j\left(T_w,T_g\right)\right] \tag{14}$$

A series of numerical experiments is conducted to investigate the effect of wall temperature (T_w) to the radiative properties of the gas mixture. The wall temperature varies from 300 K to 1500 K. The gas temperature is maintained at 500 K and the remaining test conditions are identical to that used from the previous test case as shown in Section 3.2.2. A summary of test case 2 is provided in Table 1.

Fig. 5 shows the emissivity (a) and the absorptivity (b) of the gas mixture along the line (x, 1 m, 4 m) on the top surface for different wall temperatures (300,500,1000, or 1500 K) with the gas temperature maintained at 500 K. For gas emission, the approximate approach over-predicts the emissivity by approximately 8 % to 37 %. For mixture absorptivity, the results show that absorptivity can vary significantly depending on the wall temperature. In Fig. 5b, it can be seen that the mixture absorptivity is decreasing with increasing wall temperature. When the wall temperature is identical to the gas temperature, the emissivity equals the absorptivity. This observation agrees well with the findings in [8, 10]. However, since the approximate approach does not account for the wall temperature in the determination of the radiative properties, the absorptivity is identical

Tam, Wai Cheong; Yuen, Walter.
"Assessment of Radiation Solvers of Fire Simulation Models Using RADNNET-ZM."
Paper presented at 11th Asia-Oceania Symposium on Fire Science and Technology (AOSFST), Taipei, Taiwan Province of China. October 21, 2018 - October 25, 2018.

to the emissivity. As shown in Fig. 5b, the maximum discrepancy associated with the absorptivity predicted by the approximate approach can be different from that predicted by RADNNET-ZM by more than 100 %.

Test cases	Wall conditions		Gas medium conditions		
	Temperature	Emissivity	Temperature	Composition	Pressure
Case 1	300 K	1	1000 K	Water vapor	1 atm
Case 2	300/500/1000/1500 K	1	500 K	Water vapor	1 atm

Table 1. Test conditions for different test cases.

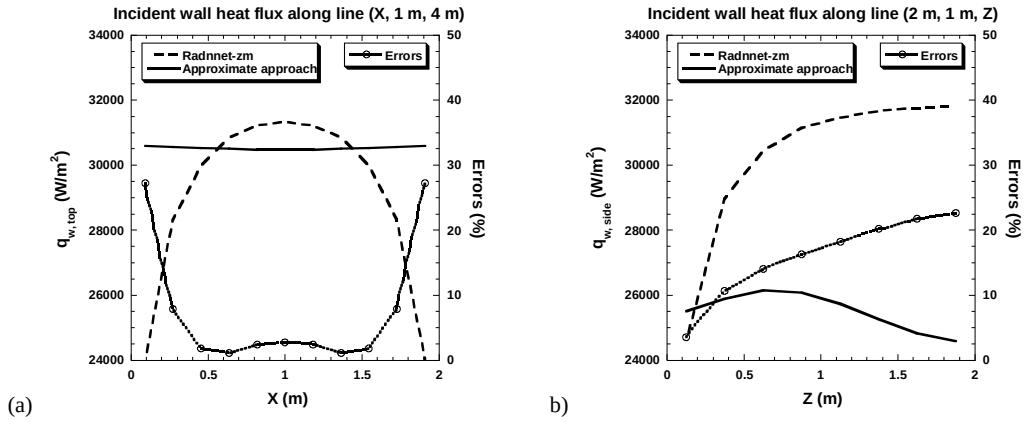


Fig. 4. Incident wall heat flux alone lines (a) [x, 1 m, 4 m] and (b) [2 m, 1 m, z] obtained from RADNNET-ZM and the approximate approach together with the error associated with the approximate approach.

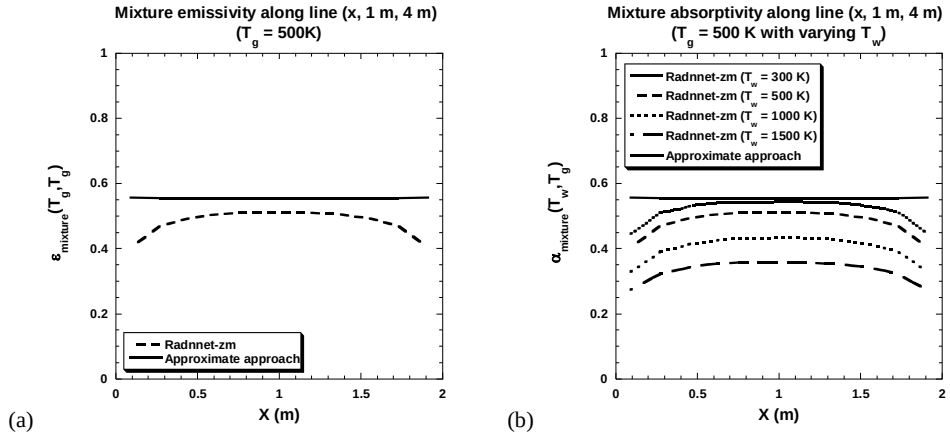


Fig. 5. (a) Mixture emissivity [$T_w = T_g$] and (b) mixture absorptivity [$T_w \neq T_g$] obtained from RADNNET-ZM and the approximate approach.

4. Conclusions

The mathematical formulation of the neural network based generalized zonal method (RADNNET-ZM) is presented. Using the concept of generic exchange factor and the method of superposition, the radiative heat transfer to bounding surfaces in an arbitrary Cartesian enclosure with an isothermal, homogeneous, non-gray medium can be evaluated.

For an isothermal medium with pure water vapor emitting at 1000 K, errors associated with RADNNET-ZM predictions are within 4 %. For the approximate approach that is used in CFAST, the errors associated with the predictions are within 28 %.

The effect of wall temperature on the evaluation of the radiative properties has been investigated. Mixture emissivity is generally not equal to mixture absorptivity. Given a fixed gaseous concentration, it can be shown while emissivity is a strong function of the gas temperature, absorptivity varies significantly for both wall and gas (absorbing) temperature. Results show

Tam, Wai Cheong; Yuen, Walter.
"Assessment of Radiation Solvers of Fire Simulation Models Using RADNNET-ZM."
Paper presented at 11th Asia-Oceania Symposium on Fire Science and Technology (AOSFST), Taipei, Taiwan Province of China. October 21, 2018 - October 25, 2018.

that the use of the Hottel's emissivity chart is ineffective and highly inaccurate for the determination of the radiative properties in an enclosure with non-gray gases. Comparing the results generated by RADNNET-ZM, the approximate approach over-predicts the emissivity by approximately 8 % to 37 %. For absorptivity, the discrepancy between two radiation solvers can be more than 100 %.

Even though the current work focuses on isothermal homogenous media, RADNNET-ZM can be readily used to accurately simulate the radiative absorption effect for different mixtures involving other species (i.e., H_2O, CO_2, and soot particulate). Solutions generated by RADNNET-ZM can be used as benchmark results to verify other radiation solvers from other commercial/CFD codes, such as FDS. For non-isothermal and in-homogeneous conditions, the concept of mean temperatures as referenced in [18] can be implemented to expand RADNNET-ZM to account for the radiative heat transfer effect in an enclosure with non-isothermal, inhomogeneous, non-gray media. This work is currently underway and results will be presented in future publications.

Acknowledgements

The authors would like to thank Kevin B. McGrattan for his constructive comments and valuable suggestions to this manuscript.

References

[1] Hottel, H. C., & Sarofim, A. F, 1967. Radiative transfer. McGraw-Hill.
[2] Fiveland, W. A., 1987. "Discrete ordinate methods for radiative heat transfer in isotropically and anisotropically scattering media." Journal of Heat Transfer 109, no. 3, pp. 809-812.
[3] Carvalho, M., Farias, T. and Fontes, P., 1991. "Predicting radiative heat transfer in absorbing, emitting, and scattering media using the discrete transfer method." Fundamentals of radiation heat transfer 160, no. 1, pp. 17-26.
[4] Grosshandler, W. L. RADCAL: A narrow-band model for radiation calculations in a combustion environment. Gaithersburg, MD: National Institute of Standards and Technology, 1993.
[5] Mazumder, S., & Modest, M. F., 2002. "Application of the full spectrum correlated-k distribution approach to modeling non-gray radiation in combustion gases." Combustion and Flame, 129(4), pp. 416-438.
[6] Cumber, P. S., Fairweather, M., & Ledin, H. S., 1998. "Application of wide band radiation models to non-homogeneous combustion systems." International Journal of Heat and Mass Transfer, 41(11), pp. 1573-1584.
[7] Choi, C. E., & Baek, S. W., 1996. "Numerical analysis of a spray combustion with nongray radiation using weighted sum of gray gases model." Combustion science and technology, 115(4-6), pp. 297-315.
[8] Yuen, W. W., Tam, W. C., & Chow, W. K., 2014. "Assessment of radiative heat transfer characteristics of a combustion mixture in a three-dimensional enclosure using RAD-NETT (with application to a fire resistance test furnace)." International Journal of Heat and Mass Transfer, 68, pp. 383-390.
[9] Yuen, W. W., 2009. "RAD-NNET, a neural network based correlation developed for a realistic simulation of the non-gray radiative heat transfer effect in three-dimensional gas-particle mixtures." International Journal of Heat and Mass Transfer 52, no. 13, pp. 3159-3168.
[10] Tam, W. C. Analysis of heat transfer in a building structure accounting for the realistic effect of thermal radiation heat transfer. Ph.D. Thesis, the Hong Kong Polytechnic University, Hong Kong, China, 2013.
[11] Peacock, R. D., McGrattan, K. B., Forney, G. P., & Reneke, P. A. CFAST–Consolidated Fire And Smoke Transport (Version 7) Volume 1: Technical Reference Guide. Technical Note, NIST, Gaithersburg, Maryland, 1, 69-71, 2015.
[12] McGrattan, K., Hostikka, S., McDermott, R., Floyd, J., Weinschenk, C., & Overholt, K. Fire dynamics simulator technical reference guide volume 1: mathematical model. NIST special publication, 1018(1), 175, 2013.
[13] Fluent, A. N. S. Y. S. Ansys fluent theory guide. ANSYS Inc., USA, 15317, 724-746, 2011.
[14] Yuen, W. W., and Takara, Ezra E., 1997. "The zonal method: A practical solution method for radiative transfer in nonisothermal inhomogeneous media." Annual review of heat transfer 8, no. 8.
[15] Yuen, W. W. & Tam, W. C. "RADNNET-ZM – The generalized zonal method for radiative transfer in multi-dimensional non-gray media." In preparation.
[16] Liu, Fengshan, 1999. "Numerical solutions of three-dimensional non-gray gas radiative transfer using the statistical narrow-band model." Journal of heat transfer 121, no. 1, pp. 200-203.
[17] Coelho, P. J., 2002. "Numerical simulation of radiative heat transfer from non-gray gases in three-dimensional enclosures." Journal of Quantitative Spectroscopy and Radiative Transfer 74, no. 3, pp. 307-328.
[18] Yuen, W.W., 2014. "Development of the concept of mean temperatures in the analysis of radiative heat transfer in an inhomogeneous non-isothermal non-gray medium." International Journal of Heat and Mass Transfer, 68, pp.259-268.

Tam, Wai Cheong; Yuen, Walter.
"Assessment of Radiation Solvers of Fire Simulation Models Using RADNNET-ZM."
Paper presented at 11th Asia-Oceania Symposium on Fire Science and Technology (AOSFST), Taipei, Taiwan Province of China. October 21, 2018 - October 25, 2018.

Life-Cycle Cost of Manufactured Goods: A Case Study in US Ground Passenger Transportation

Douglas Thomas, Economist

National Institute of Standards and Technology

Disclaimer: Certain trade names and company products are mentioned in the text in order to adequately specify the technical procedures and equipment used. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products are necessarily the best available for the purpose.

Abstract: National governments invest in research and development to advance efficiency and spur economic growth. There are, however, few studies that identify where investments will have the largest possible return on investment. This lack of research can result in the funding of investments with suboptimal returns. Initial research in this area has focused on examining production costs; however, to identify high-return areas of research, efforts need to be taken further to include both the production and use of a product. This paper examines the life-cycle cost of passenger ground transportation as a proof of concept to identify those items that have both a high cost and high environmental impact. Public research that focuses on these items has the potential to be more economical than other areas. This paper uses US input-output data from the US Bureau of Economic Analysis, data from the American Time Use Survey, and environmentally extended input-output data to examine the supply chain for production and use of ground transportation equipment. This paper is unique in that it identifies the costs, some of which are not documented in GDP (i.e., uncompensated time use), along with the environmental impacts of producing and using a class of manufactured goods. The Pareto principle, which posits that roughly 80 % of a problem is due to 20 % of the causes, is utilized for targeting specific efficiency solutions. Those supply chain entities that are above the 80th percentile for both financial costs and environmental impacts are identified. The robustness of this identification is examined using Monte Carlo techniques. Forty-three supply chain entities were identified as being above the 80th percentile for cost, measured in value added, and environmental impact with six being above the 95th percentile for both.

1. Introduction

As illustrated in Figure 1, governments seek to advance efficiency in the economy by reducing inputs and negative externalities (represented in red with down arrows indicating a decrease), such as environmental impact, while increasing output and product function (represented in green with up arrows indicating an increase). The result is an increase in the quality and quantity of production at lower per unit costs and environmental impacts. These types of advancements facilitate sustained economic growth that increases average income.[1] On their own accord, firms pursue efficiency improvements that increase profit; however, there are limited incentives for a firm to pursue activities in which they cannot sufficiently capture enough of the benefit, such as environmental sustainability. Additionally, there are potential efficiency improvements that might not be achieved due to market failures. For this reason, governments invest in research and development to advance efficiency that results in sustainable economic growth. Unfortunately, there are a limited number of studies that identify the research areas that have the potential for having the highest return on investment. The result is that governments can often fund suboptimal investments.

Previous work by Thomas and Kandaswamy examined assembly-centric products (i.e, machinery, electronics, computers, and transportation equipment) to identify those supply chain points that accounted for a disproportional amount of the cost of production.[2] In another paper, Thomas and Kandaswamy

[1] Weil, David N. Economic Growth. United States: Pearson Education Inc., 2005. 181

[2] Thomas, Douglas and Anand Kandaswamy. "Identifying High Resource Consumption Areas of Assembly-Centric Manufacturing in the United States." Journal of Technology Transfer. 2017.

Thomas, Douglas.
"Life-Cycle Cost of Manufactured Goods: A Case Study in US Ground Passenger Transportation."
Paper presented at 26th International Input-Output Conference, Juiz de Fora, Brazil. June 25, 2018 - June 29, 2018.

examined, at the industry level, material flow time, which is often used by manufacturers to track and improve competitiveness.[3] The same authors used input-output analysis to identify supply-chain points that consume high levels of resources, including financial and environmental resources.[4] Each of these papers focused on the inputs and/or negative externalities associated with production. This paper extends that work by examining both production and the function of a product.

With a multitude of products, processes, and activities, a holistic approach will require a systematic method to examine production and utilization. The standard categorization of industry activity combined with input-output analysis, which was originally developed by economist Leontief,[5] provides a foundation for such an approach. Input-output models are typically used to estimate the impact of a shift in demand for a good or service, but they also provide information on inter-industry activity, making such models an invaluable resource for industry-by-industry resource use within the US economy.

A frequently invoked axiom posits that roughly 80 % of a problem can be traced to 20 % of the cause(s), a phenomenon referred to as the Pareto principle.[6] This paper identifies those cost items that account for a disproportionally high level of resource consumption compared to other cost items. A method is developed and used to examine US ground passenger transportation as a case study. Passenger transport represents multiple industries each with many supply chain costs. A multi-factor approach is used to measure environmental impact and value added. Note that value added is the revenue that an establishment receives less the purchases from other establishments; thus, it is the establishment's contribution to the cost that the consumer bears in purchasing the final product.

The purpose of this paper is to facilitate the identification of economy-wide opportunities for researched efficiency improvements in the production and use of products. Researchers are unable to identify and compare all potential research topics that impact efficiency; thus, a method is needed to create a pool of high return investments to select from. Return on investment (ROI) can be represented as:

$$ROI = \frac{Benefit\ of\ Investment - cost\ of\ Investment}{Cost\ of\ Investment}$$

Cost categories, which in this paper are classified as NAICS codes, represent similar activities occurring within one location. Additionally, a high cost, as measured in value added, primarily occurs either because of high cost processes (through labor or profit) or a high volume of production (i.e., many units); thus, an efficiency improvement in a high cost area is, likely, to have a greater potential benefit through spillover than an efficiency improvement in a low-cost area. Stated another way, high cost categories represent a potentially target rich environment of investments with a high level of benefits that can result in a high return on investment. Additionally, those production activities that have both a high cost and high environmental impact, which will be referred to as consuming a high level of resources, provide a robust opportunity for efficiency improvement affecting multiple stakeholders (i.e., citizens, consumers, and producers). Public entities, trade organizations, and other change agents that seek to maximize efficiency improvement through innovative solutions can search for potential research areas in high cost areas, increasing the likelihood of identifying high return investments. After identifying high cost/impact areas, potential research topics within those cost areas can be identified and compared. It is important to

[3] Thomas, Douglas and Anand Kandaswamy. 2017 "An Examination of National Supply-Chain Flow Time." Economic Systems Research. http://www.tandfonline.com/doi/full/10.1080/09535314.2017.1407296

[4] Thomas, Douglas and Anand Kandaswamy. 2017

[5] Miller, Ronald E. and Peter D. Blair. Input-Output Analysis: Foundations and Extensions. Second Edition. New York: Cambridge University Press, 2009.

[6] Hopp, Wallace J. and Mark L. Spearman. Factory Physics. Third Edition. Long Grove, IL: Waveland Press, 2008. 674.

note that there are a number of factors that are relevant to choosing the most economical investments to improve efficiency. The approach in this paper is a method for examining one of those factors.

2. Methods

This paper uses input-output data and analysis to examine various aspects of the supply chain for ground passenger transportation. Business and personal expenditures on transportation are inserted into an input-output model, which estimates value added and environmental impact by industry. This is combined with an estimate of personal time use for transportation. Industries are categories of establishments (i.e., physical locations of economic activity) based on the product being produced and the processes being used. Commodities (i.e., products and services) are exchanged between industries and are also delivered to the final consumer. For a particular finished commodity, the value added from each industry and the environmental impact from each industry is estimated. The methods in this paper build on those in Thomas and Kandaswamy (2016)[7], Thomas and Kandaswamy (2016)[8], and Thomas and Kneifel (2016).[9] The analysis is examining the life-cycle cost and environmental impact of ground passenger transportation (i.e., cost and environmental impact for production, energy for use, maintenance, repair, and time use), but will look at one year to estimate it. At the national level, expenditures have relatively similar expenditures year to year. This analysis assumes that expenditures remain the same; therefore, regardless of whether one examines one year or thirty years, the relative ranking of the costs will remain the same.

2.1. Environmental Impact

The measure of environmental impact is calculated using input-output analysis combined with TRACI 2 impact categories and the Analytical Hierarchy Process to weight the categories. A description of the calculations is below.

Input-Output Analysis: The Make-Use tables are used for Input-Output analysis.[10] The model operates under constant returns to scale and thus ignores potential economies of scale.[11] The model also assumes that a sector uses inputs in fixed proportions. These issues are, typically, relevant to analyses that examine the impact of a change in demand.[12] This paper is not seeking to predict the impact of a change in demand, but rather seeks to track the total resources used for the production of particular goods; therefore, ignoring economies of scale and assuming sectors use inputs in fixed proportions has minimal impact on this analysis. This paper also uses an industry-by-commodity Input-Output format as outlined in Horowitz

[7] Thomas, Douglas and Anand Kandaswamy. "Identifying High Resource Consumption Areas of Assembly-Centric manufacturing in the United States." National Institute of Standards and Technology. White paper. 2016.
[8] Thomas, Douglas and Anand Kandaswamy. "Improving Manufacturing Efficiency through Supply-Chain Flow Time." National Institute of Standards and Technology. White paper. 2016.
[9] Thomas, Douglas and Joshua Kneifel. "Identifying Environmental Impact Hotspots of Assembly-Centric Manufacturing in the US." National Institute of Standards and Technology. White paper. 2016.
[10] Miller, Ronald E. and Peter D. Blair. Input-Output Analysis: Foundations and Extensions. Second Edition. New York: Cambridge University Press, 2009. 135-138.
[11] Miller. 16.
[12] Horowitz, Karen J. and Mark A. Planting. Concepts and Methods of the US Input-Output Accounts. Bureau of Economic Analysis. September 2006. http://www.bea.gov/papers/pdf/IOmanual_092906.pdf

and Planting (2006), which accounts for the fact that an industry may produce more than one commodity or product, such as secondary products and by-products.[13, 14, 15]

An input-output analysis develops a total requirements matrix that when multiplied by the vector of final demands equals the output needed for production. The total requirements matrix is developed using the methods outlined in Horowitz and Planting (2006):

Equation 1

$$X = W(I - BW)^{-1} * Y$$

Where:

$X =$ Vector of output required to produce final demand

$Y =$ Vector of final demand

$W = (I - \hat{p})D$

$B = U\hat{g}^{-1}$

$I =$ Identity matrix

$D = V\hat{q}^{-1}$

$p =$ A column vector in which each entry shows the ratio of the value of scrap produced in each industry to the industry's total output.

$U =$ Intermediate portion of the use matrix in which the column shows for a given industry the amount of each commodity it uses—including noncomparable imports, scrap, and used and secondhand goods. This is a commodity-by-industry matrix.

$V =$ Make matrix, in which the column shows for a given commodity the amount produced in each industry. This is an industry-by-commodity matrix. V has columns showing only zero entries for noncomparable imports and for scrap.

$g =$ A column vector in which each entry shows the total amount of each industry's output, including its production of scrap. It is an industry-by-one vector.

[13] Ibid.
[14] Miller. 184.
[15] European Commission. Eurostat Manual of Supply, Use, and Input-Output Tables. 2008 Edition. 2008. Accessed September 2016. http://ec.europa.eu/eurostat/documents/3859598/5902113/KS-RA-07-013-EN.PDF/b0b3d71e-3930-4442-94be-70b36cea9b39?version=1.0.

q = A column vector in which each entry shows the total amount of the output of a commodity. It is a commodity-by-one vector.

$\widehat{\ }$ A symbol that when placed over a vector indicates a square matrix in which the elements of the vector appear on the main diagonal and zeros elsewhere.

In Equation 1, a total requirements matrix $W(I - BW)^{-1}$ is multiplied by a vector of final demand for commodities Y to estimate the total output X. All variables in Equation 1 have known values in the input output data. The output X required to produce an alternate level of final demand can be calculated by altering the final demand vector from the actual final demand Y in the input output data to Y'. For this analysis, Y' has the actual final demand for assembly-centric commodities and zero for other commodities. This alteration reveals the output needed to produce only assembly-centric commodities.

Environmental Impact Categories: The TRACI 2 impact categories are each an aggregation of multiple emissions converted to a common physical unit. For example, the global warming impact category includes impacts of many pollutants, such as carbon dioxide (CO_2), methane (CH_4), nitrous oxide (NO_X), and fluorinated gases, which are converted to their carbon dioxide equivalent (CO_2e) impact and aggregated to estimate the total impact for that impact category. The environmental impacts are measured in terms of the common physical unit per dollar of output. The impact can be calculated by multiplying the output in the Input-Output analysis by the impact categories.

Impact Category Weights: Having 12 environmental impact categories makes it difficult to rank industry environmental activity; therefore, the 12 impact categories have been combined into a single environmental metric using the Analytical Hierarchy Process (AHP). AHP is a mathematical method for developing weights using normalized eigenvalues. It involves making pairwise comparisons of competing items based on a multilevel hierarchy developed by the user. The weights used in this paper were developed for the BEES software and can be seen in Table 1.[16] This paper uses 12 of the 13 impact categories for which weights were developed. Indoor Air Quality (IAQ) is excluded because it is more applicable to the design of buildings and ventilation systems rather than to manufacturing activities. The weight of IAQ is proportionally allocated to the other 12 impact categories. The final metric for each industry or industry/commodity combination is the proportion of the total impact from assembly-centric products. The percent of environmental impacts, based on the weights, are calculated using the following equation:

[16] Lippiatt, Barbara, Anne Landfield Greig, and Priya Lavappa. Building for Environmental and Economic Sustainability. National Institute of Standards and Technology. 2010. Accessed September 2016. http://www.nist.gov/el/economics/BEESSoftware.cfm.

Equation 2

$$Env_{z,Y'} = \frac{x_{z,Y'} * GWP_z}{\sum_{i=1}^{n} x_{i,Y'} * GWP_i} * 0.30 + \frac{x_{z,Y'} * Acid_z}{\sum_{i=1}^{n} x_{i,Y'} * Acid_i} * 0.03 + \frac{x_{z,Y'} * HHA_z}{\sum_{i=1}^{n} x_{i,Y'} * HHA_i} * 0.09$$
$$+ \frac{x_{z,Y'} * Eut_z}{\sum_{i=1}^{n} x_{i,Y'} * Eut_i} * 0.06 + \frac{x_{z,Y'} * OD_z}{\sum_{i=1}^{n} x_{i,Y'} * OD_i} * 0.02 + \frac{x_{z,Y'} * Sm_z}{\sum_{i=1}^{n} x_{i,Y'} * Sm_i} * 0.04$$
$$+ \frac{x_{z,Y'} * Eco_z}{\sum_{i=1}^{n} x_{i,Y'} * Eco_i} * 0.07 + \frac{x_{z,Y'} * HHC_z}{\sum_{i=1}^{n} x_{i,Y'} * HHC_i} * 0.08 + \frac{x_{z,Y'} * HHNC_z}{\sum_{i=1}^{n} x_{i,Y'} * HHNC_i} * 0.05$$
$$+ \frac{x_{z,Y'} * PE_z}{\sum_{i=1}^{n} x_{i,Y'} * PE_i} * 0.10 + \frac{x_{z,Y'} * LU_z}{\sum_{i=1}^{n} x_{i,Y'} * LU_i} * 0.06 + \frac{x_{z,Y'} * WC_z}{\sum_{i=1}^{n} x_{i,Y'} * WC_i} * 0.08$$

Where

$Env_{z,Y'} =$ Environmental impact from industry z for final demand Y'

$GWP_z =$ Global warming potential per dollar of output for industry z

$Acid_z =$ Acidification per dollar of output for industry z

$HHA_z =$ Human health –criteria air pollutants – per dollar of output for industry z

$Eut_z =$ Eutrophication per dollar of output for industry z

$OD_z =$ Ozone depletion per dollar of output for industry z

$Sm_z =$ Smog per dollar of output for industry z

$Eco_z =$ Ecotoxicity per dollar of output for industry z

$HHC_z =$ Human health – carcinogens – per dollar of output for industry z

$HHNC_z =$ Human health – non-carcinogen – per dollar of output for industry z

$PE_z =$ Primary energy consumption per dollar of output for industry z

$LU_z =$ Land use per dollar of output for industry z

$WC_z =$ Water consumption per dollar of output for industry z

$x_{z,Y'} =$ Output for industry z with final demand Y'

i = industry i through n

2.2. Value Added

The total requirements matrix $W(I - BW)^{-1}$ from Equation 1, which shows the total output required to meet a given level of final demand, is multiplied by final demand in the input-output data to estimate the total output. The output required to produce a particular level of final demand can be calculated by altering final demand to Y'. For this analysis, Y' equals final demand for those NAICS codes representing the production and use of ground passenger transportation equipment and zero for those that do not.

Value added is calculated by assuming the proportion of output needed to produce a commodity is the same proportion of value added, which is consistent with methods proposed by Miller (2009). The proportions calculated using the input-output analysis are then multiplied by the value added and scaled to 2014 dollars using the estimate of gross output for that year:

Equation 3

$$VA_{z,Y',2014} = \frac{x_{z,Y',2007}}{x_{z,2007}} * VA_{z,2007} * \left(\frac{x_{z,2014}}{x_{z,2007}}\right)$$

Where

$VA_{z,Y',2014}$ = Value added from industry z with final demand Y' in 2014

$x_{z,2007}$ = Total output for industry z in 2007

$x_{z,2014}$ = Total output for industry z in 2014

$x_{z,Y',2007}$ = Output for industry z with final demand Y' in 2007

$VA_{z,2007}$ = Total value added from industry z in 2007

Imports are calculated in a similar fashion, where the proportion of total output used from a particular industry is the same for imports.

2.3. Production and Use

To examine both the production and use of a product, the Bureau of Economic Analysis' data on personal consumption expenditures by function is used, which includes categories for transportation. A list of resource consumption categories is provided in Figure 2. It includes items that are documented in the nation's gross domestic product and items that are not included. Some passenger transport is purchased by consumers while other purchases are made by other industries. Figure 2 separates these categories. Both categories have to purchase or pay for fuels along with maintenance and repair. Some of these items are imported while others are produced domestically. Consumers face costs, however, that are not documented in the economy, including the time spent in transport and time spent on maintenance and repair that they themselves conduct. There are also infrastructure costs that the public sector bears. Unfortunately, the environmental impact of personal fuel consumption is not captured in this assessment.

The US Census Bureau's American Time Use Survey is used to measure uncompensated labor such as driving time. Each purchase, including that for vehicles, public transport, vehicle maintenance and repair, and fuel, is entered into the input-output model. This calculation assumes that imported items face similar costs and impact as those produced domestically. Although this is not strictly accurate, imported products have similar materials and components. The value of uncompensated time is calculated by multiplying the average time use per year by the average hourly compensation for 2014, which is $32.05.

2.4. Sensitivity Analysis

This analysis uses data from previous years to guide research decisions for current industry activity, which results in some uncertainty. In order to account for this uncertainty, a probabilistic sensitivity analysis was conducted using Monte Carlo analysis. Examinations of uncertainty in environmental Input-

Output analysis have used both fuzzy set theory and stochastic models[17, 18, 19, 20, 21]; however, with there being limited in-depth examinations of uncertainty, there is not a consensus on a specific approach.[22] Monte Carlo analysis is based on works by McKay, Conover, and Beckman[23] and by Harris[24] that involves a method of model sampling. Monte Carlo simulation methods are superior to deterministic modeling for our purposes because deterministic modeling uses single-point estimates while Monte Carlo generates a probability distribution for every single variable of interest and allows for a comprehensive comparison of those probabilities.

The method was implemented using the Crystal Ball software product[25], a software add-in for spreadsheets. Specification involves defining which variables are to be simulated, the distribution of each of these variables, and the number of iterations performed. The software then randomly samples from the probabilities for each input variable of interest.

For this analysis, the industries that are above the 80th percentile for both environmental impact and value added were included in the Monte Carlo analysis, which includes 43 industries. For the environmental impact, each of the TRACI factors were varied by +/- 10 % and the weights are varied by +/- 25 %. For value added, each industry was varied by +/- 25 %. The remaining industries are varied together by +/- 10 %. Each variation uses a triangular distribution where the base case is the most likely value. Although different levels of variation could be selected, it has been shown that this level of error is consistent with previous works.[26, 27] This simulation contained 10 000 iterations.

[17] Raina, Roma, Mini Thomas. Fuzzy vs. Probabilistic Techniques to Address Uncertainty for Radial Distribution Load Flow Simulation. Energy and Power Engineering. 2012 (4) 99-105. http://dx.doi.org/10.4236/epe.2012.42014

[18] Egilmez, Gokhan, Serkan Gumus, Murat Kucukvar, Omer Tatari. "A Fuzzy Data Envelopment Analysis Framework for Dealing with Uncertainty Impacts of Input-Output Life Cycle Assessment Models on Eco-efficiency Assessment." Journal of Cleaner Production. 2016. doi: 10.1016/j.jclepro.2016.03.111

[19] Beynon, Malcolm James and Max Munday. An Aggregated Regional Economic Input-Output Analysis within a Fuzzy Environment. Spatial Economic Analysis. November 2007. 2 (3) 281-296.

[20] Beynon, Malcolm J, Max Munday, and Annette Roberts. Ranking Sectors using Fuzzy Output Multipliers. Economic Systems Research. 2005. 17 (3) 237-253.

[21] Temurshoev, Umed. "Uncertainty Treatment in Input-Output Analysis." 2015. http://loyolaandnews.es/loyolaecon/wp-content/uploads/2015/12/Uncertainty-treatment-in-Input-Output-analysis.pdf

[22] Diaz, Barbara and Antonio Morillas. Incorporating Uncertainty in the Coefficients and Multipliers of an IO Table: A Case Study. Papers in Regional Science. 90 (4) 845-861.

[23] McKay, M. C., W. H. Conover, and R.J. Beckman, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," 1979, Technometrics (Vol. 21): pp. 239-245.

[24] Harris, Carl M, Issues in Sensitivity and Statistical Analysis of Large-Scale, Computer-Based Models, NBS GCR 84-466, Gaithersburg, MD: National Bureau of Standards, 1984.

[25] Crystal Ball, Crystal Ball 11.1.2.3 User Manual. Denver, CO: Decisioneering, Inc. 2013.

[26] European Science and Technology Observatory. Environmental Impact of Products: Analysis of the Life Cycle Environmental Impacts Related to the Final Consumption of the EU-25. 2006. http://ec.europa.eu/environment/ipp/pdf/eipro_report.pdf

[27] Temurshoev, Umed. "Uncertainty Treatment in Input-Output Analysis." 2015. http://loyolaandnews.es/loyolaecon/wp-content/uploads/2015/12/Uncertainty-treatment-in-Input-Output-analysis.pdf

3. Data

Three datasets are needed to examine costs and environmental impacts. These datasets include the Bureau of Economic Analysis (BEA) Benchmark Input-Output data, Personal Consumption Expenditures from the BEA, environmentally extended input-output data, and the American Time Use Survey.

Input-Output Data and Personal Consumption Expenditures: Every five years the BEA computes benchmark input-output tables, which tends to have over 350 industries.[28] The data is provided in the form of make and use tables, with their corresponding matrices replacing the Leontief method.[29] In the US, industries are categorized by NAICS codes. There are two types of make and use tables: "standard" and "supplementary." Standard tables closely follow NAICS and are consistent with other economic accounts and industry statistics, which classify data based on establishment. Note that in this context an "establishment" is a single physical location where business is conducted. This should not be confused with an "enterprise" such as a company, corporation, or institution. Establishments are classified into industries based on the primary activity within the NAICS code definitions; however, establishments often have multiple activities. An establishment is classified based on its primary activity. Data for an industry reflects all the products made by the establishments within that industry; therefore, secondary products are included. Supplementary make-use tables reassign secondary products to the industry in which they are primary products. The data in this report utilizes the standard make-use tables. The BEA uses the data for the input-output accounts to also estimate personal consumption expenditures.*Environmental Data*: For environmental data, this paper applies a suite of environmentally extended Input-Output databases for Life Cycle Assessments (LCA) developed under contract for NIST by Dr. Sangwon Suh of the Bren School of Environmental Science and Management at the University of California, Santa Barbara.[30] This data has been utilized in a number of environmental efforts, including NIST's Building for Environmental and Economic Sustainability (BEES)[31] and Building Industry Reporting and Design for Sustainability (BIRDS)[32] software and related publications. This data utilizes the 12 TRACI 2 impact categories: global warming potential, primary energy consumption, human health – criteria air pollutants, human health – carcinogens, water consumption, ecological toxicity[33], eutrophication[34], land use, human health – non-carcinogens, smog formation, acidification, and ozone depletion. The units of measurement are provided in Table 1. This environmental data is organized by 2002 BEA codes for the Benchmark Input-Output tables, and matched and adjusted to the 2007 BEA Input-Output tables. The environmental data was adjusted from being in impact units per 2002 dollars to impact units per 2007 dollars using the consumer price index from the Bureau of Labor Statistics.

American Time Use Survey: The American Time Use Survey estimates how and where the US population spends its time, including time spent at work, leisure activities, childcare, transportation, and household activities. The data is collected from a sample size of approximately 40 500 and conducted annually.

---

[28] Bureau of Economic Analysis. Input-Output Accounts Data. November 2014. Accessed September 2016. http://www.bea.gov/industry/io_annual.htm.

[29] A System of National Accounts, Studies in Methods, Series F/No. 2/Rev. 3, New York, United Nations, 1968.

[30] This work is based on Suh, S. Developing a sectoral environmental database for input-output analysis: the comprehensive environmental data archive of the US, Economic Systems Research. 17: 4(2005): 449-469.

[31] National Institute of Standards and Technology. Building for Environmental and Economic Sustainability. Accessed September 2016. http://www.nist.gov/el/economics/BEESSoftware.cfm.

[32] National Institute of Standards and Technology. Building Industry Reporting and Design for Sustainability. Accessed September 2016. https://birdscom.nist.gov/.

[33] The potential of a chemical released into the environment to harm terrestrial and aquatic ecosystems.

[34] The addition of mineral nutrients to the soil or water, which in large quantities can result in generally undesirable shifts in the number of species in ecosystems and a reduction in ecological diversity

4. Results and Discussion

Public entities, trade organizations, and other change agents that seek to maximize efficiency improvement through innovative solutions must prioritize their efforts to get the largest reduction per expenditure dollar. In a world of limited and scarce resources, it is not technically feasible to identify all possible research topics, conduct an economic analysis of each, and identify those with the highest return. Rather, researchers can only identify a selection of the possible R&D topics. Those topics that are within high cost areas of production and use have a higher likelihood of having a large impact and return-on-investment. After identifying high cost/impact areas, potential research topics within those cost areas can be identified and compared.

Table 2 presents the costs associated with ground passenger transport along with an estimate of the uncompensated time spent in transport, referred to as resources. The total is approximately $4.9 trillion, which is 14.9 % of the total resources in the US (the total includes an estimate for uncompensated time spent working). The largest cost item is the consumer transport time, which amounts to $3.3 trillion or 68 % of the total resource cost. A great deal of travel time is spent going to and from the store to purchase goods/services, as seen in Table 3. The next largest is transportation to and from work with leisure/other being third. The second largest cost in Table 2 is the fuels, maintenance, and repair. The consumer, commercial, and industrial purchases amount to $1.1 trillion. The implication of these results is that the product design, infrastructure, and reducing the need for transportation can have a disproportional impact, compared to other cost items, on the resource consumption for ground passenger transit. For instance, reducing the need for transportation by allowing employees to telecommute and improving telecommunications might have a larger reduction in resource consumption than other targeted efforts, as it reduces the largest cost item – transport time; however, only a limited number of jobs can facilitate telecommuting. Other potential resource saving efforts might include alleviating traffic congestion or facilitating the delivery of goods. Increasing fuel efficiency and reducing maintenance/repair needs can also have a disproportional impact. The advancement of autonomous (i.e., self-driving) vehicles can also improve efficiency, as it could, potentially, allow the operator to conduct other activities.

The items included as expenditures (i.e., everything except uncompensated time) were examined using input-output analysis to identify other top cost items. Forty-three industries were identified as being above the 80th percentile for both value added and environmental impact, as illustrated in Figure 3 and listed in Table 4. This figure and table includes the resources used for both production and use of transportation equipment. Six industries are listed as having both the environmental impact and value added as being above the 95th percentile: “211000 oil and gas extraction”; “iron and steel mills and ferroalloy manufacturing”; “336111 automobile manufacturing”; “324111 automobile manufacturing”; “324110 petroleum refineries”; “485000 transit and ground passenger transportation”; and “811100 automotive repair and maintenance.” It is important to note that industries associated with fuel production (e.g., “211000 oil and gas extraction” and “324110 petroleum refineries”) are associated with even greater environmental impact, as this analysis was unable to capture the burning of fuel in personal vehicles. Efficiency improvements in these areas can have a disproportional impact on resource consumption when compared to other supply chain industries (i.e., cost items). For instance, a 1 % reduction in “211000 oil and gas extraction” amounts to a total 0.35 % reduction in the total cost, a larger reduction than any other industry. It also would result in a 0.16 % decrease in the environmental impact, which is the second largest of any supply chain industry. Increasing fuel efficiency through light weighting can reduce costs and environmental impacts from “211000 oil and gas extraction” and “324110 petroleum refineries” while also reducing material costs. Increased efficiency in automobile assembly can reduce “336111 automobile manufacturing.” Currently, it is difficult for consumers to compare the maintenance costs of various vehicles; therefore, standards for measuring and comparing maintenance and repair costs for automobiles might facilitate reducing costs from “811100 automotive repair and maintenance.”

The value added items and the environmental impacts shown in Figure 3 correlate with a coefficient of 0.78, suggesting that production costs and environmental impacts can be reduced simultaneously. The results from the Monte Carlo analysis, also shown in Table 4, do not have any industry above the 80th percentile that varies more than 5.2 percentile points. Only four of the industries drop below the 80th percentile. Moreover, these results suggest that the rankings of cost and environmental impact are fairly robust.

5. Conclusion

National governments invest in research and development efforts that advance efficiency and spur economic growth. There are, however, few studies that identify the efforts that will have the largest possible return on investment, resulting in the funding of investments with suboptimal returns. Previous work has focused on examining production costs; however, to identify high-return areas of research all costs, both the production and use of a product, need to be considered. This paper examines the life-cycle cost of passenger ground transportation as a proof of concept to identify areas of public research that might have a high return on investment. It uses input-output analysis to examine the supply chain for production and use of transportation equipment. Future research might expand the analysis to examine multiple product categories rather than focusing on transportation alone.

Works Cited

Beynon, Malcolm J, Max Munday, and Annette Roberts. Ranking Sectors using Fuzzy Output Multipliers. Economic Systems Research. 2005. 17 (3) 237-253.

Beynon, Malcolm James and Max Munday. An Aggregated Regional Economic Input-Output Analysis within a Fuzzy Environment. Spatial Economic Analysis. November 2007. 2 (3) 281-296.

Crystal Ball, Crystal Ball 11.1.2.3 User Manual. Denver, CO: Decisioneering, Inc. 2013.

Diaz, Barbara and Antonio Morillas. Incorporating Uncertainty in the Coefficients and Multipliers of an IO Table: A Case Study. Papers in Regional Science. 90 (4) 845-861.

Egilmez, Gokhan, Serkan Gumus, Murat Kucukvar, Omer Tatari. "A Fuzzy Data Envelopment Analysis Framework for Dealing with Uncertainty Impacts of Input-Output Life Cycle Assessment Models on Eco-efficiency Assessment." Journal of Cleaner Production. 2016. doi: 10.1016/j.jclepro.2016.03.111

European Commission. Eurostat Manual of Supply, Use, and Input-Output Tables. 2008 Edition. 2008. Accessed September 2016. http://ec.europa.eu/eurostat/documents/3859598/5902113/KS-RA-07-013-EN.PDF/b0b3d71e-3930-4442-94be-70b36cea9b39?version=1.0.

European Science and Technology Observatory. Environmental Impact of Products: Analysis of the Life Cycle Environmental Impacts Related to the Final Consumption of the EU-25. 2006. http://ec.europa.eu/environment/ipp/pdf/eipro_report.pdf

Harris, Carl M, Issues in Sensitivity and Statistical Analysis of Large-Scale, Computer-Based Models, NBS GCR 84-466, Gaithersburg, MD: National Bureau of Standards, 1984.

Hopp, Wallace J. and Mark L. Spearman. Factory Physics. Third Edition. Long Grove, IL: Waveland Press, 2008. 674.

Horowitz, Karen J. and Mark A. Planting. Concepts and Methods of the US Input-Output Accounts. Bureau of Economic Analysis. September 2006. http://www.bea.gov/papers/pdf/IOmanual_092906.pdf

Lippiatt, Barbara, Anne Landfield Greig, and Priya Lavappa. Building for Environmental and Economic Sustainability. National Institute of Standards and Technology. 2010. Accessed September 2016. http://www.nist.gov/el/economics/BEESSoftware.cfm.

McKay, M. C., W. H. Conover, and R.J. Beckman, "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," 1979, Technometrics (Vol. 21): pp. 239-245.

Miller, Ronald E. and Peter D. Blair. Input-Output Analysis: Foundations and Extensions. Second Edition. New York: Cambridge University Press, 2009.

National Institute of Standards and Technology. Building for Environmental and Economic Sustainability. Accessed September 2016. http://www.nist.gov/el/economics/BEESSoftware.cfm.

National Institute of Standards and Technology. Building Industry Reporting and Design for Sustainability. Accessed September 2016. https://birdscom.nist.gov/.

Raina, Roma, Mini Thomas. Fuzzy vs. Probabilistic Techniques to Address Uncertainty for Radial Distribution Load Flow Simulation. Energy and Power Engineering. 2012 (4) 99-105. http://dx.doi.org/10.4236/epe.2012.42014

Suh, S. Developing a sectoral environmental database for input-output analysis: the comprehensive environmental data archive of the US, Economic Systems Research. 17: 4(2005): 449-469.

Temurshoev, Umed. "Uncertainty Treatment in Input-Output Analysis." 2015. http://loyolaandnews.es/loyolaecon/wp-content/uploads/2015/12/Uncertainty-treatment-in-Input-Output-analysis.pdf

Thomas, Douglas and Anand Kandaswamy. "Identifying High Resource Consumption Areas of Assembly-Centric manufacturing in the United States." National Institute of Standards and Technology. White paper. 2016.

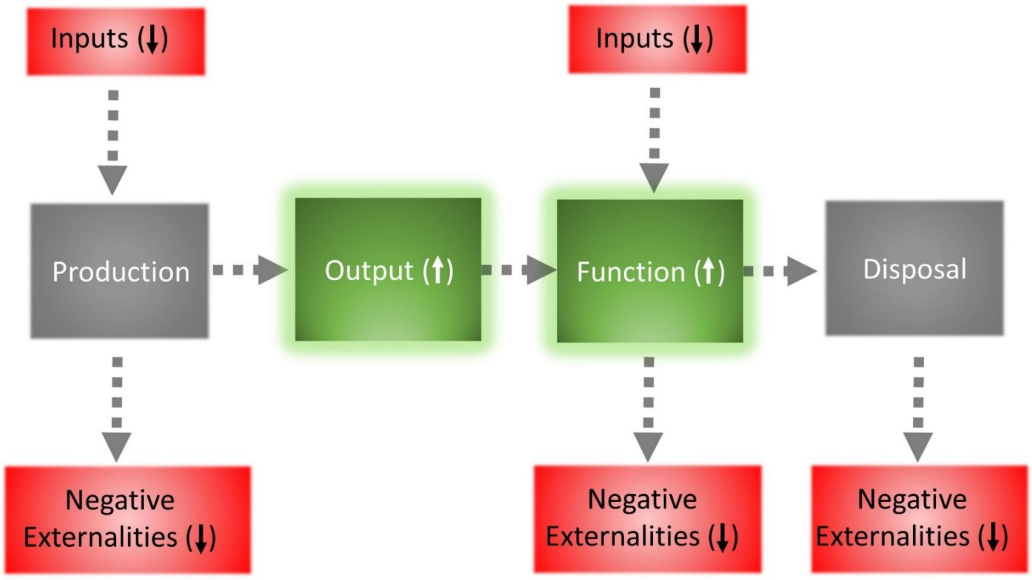
Thomas, Douglas and Anand Kandaswamy. "Improving Manufacturing Efficiency through Supply-Chain Flow Time." National Institute of Standards and Technology. White paper. 2016.

Thomas, Douglas and Anand Kandaswamy. 2017 "An Examination of National Supply-Chain Flow Time." Economic Systems Research. http://www.tandfonline.com/doi/full/10.1080/09535314.2017.1407296

Thomas, Douglas and Joshua Kneifel. "Identifying Environmental Impact Hotspots of Assembly-Centric Manufacturing in the US." National Institute of Standards and Technology. White paper. 2016.

Weil, David N. Economic Growth. United States: Pearson Education Inc., 2005. 181

Figure 1: Implicit Goals of Public Research in Manufacturing



Thomas, Douglas.
"Life-Cycle Cost of Manufactured Goods: A Case Study in US Ground Passenger Transportation."
Paper presented at 26th International Input-Output Conference, Juiz de Fora, Brazil. June 25, 2018 - June 29, 2018.

Figure 2: Domestic Resource Consumption Map for Discrete Manufactured Goods

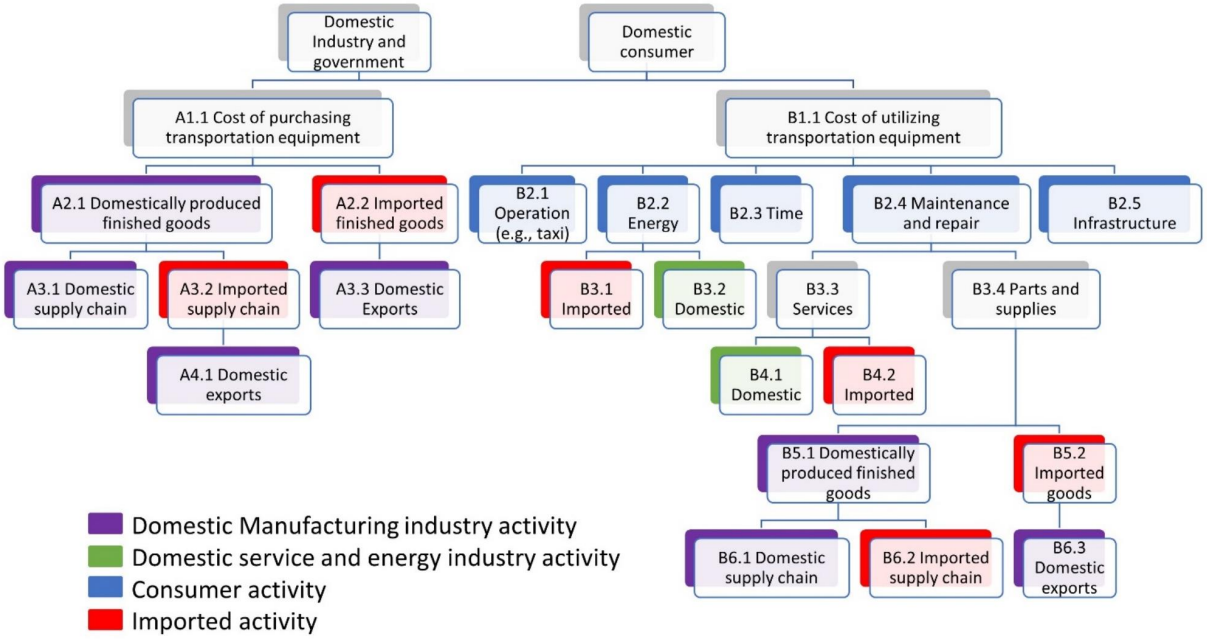


Table 1: Environmental Impact Categories and Weights for Assessing Impact

Items to be measured	**Units**	**Weights**
Global Warming Potential	kg CO_2 eq	0.30
Acidification	H^+ moles eq	0.03
Human Health- Criteria Air Pollutants	kg PM_{10} eq	0.09
Eutrophication	kg N eq	0.06
Ozone Depletion	kg CFC-11 eq	0.02
Smog	kg O_3 eq	0.04
Ecotoxicity	CTUe	0.07
Human Health - Carcinogens	CTUHcan	0.08
Human Health – Non- Carcinogens	CTUHnoncan	0.05
Primary Energy Consumption	thousand BTU	0.10
Land Use	acre	0.06
Water Consumption	kg	0.08

Table 2: Resources Related to Ground Passenger Transport, 2014

	$Billion 2014	Percent of Total Resources
Consumer purchases (A1.1)	**266.1**	**0.8%**
New motor vehicles	266.1	0.8%
Consumer maintenance, repair, and energy (B2.1, B2.2, B2.4)	**931.8**	**2.8%**
Public Transportation	99.8	0.3%
Motor vehicle parts and accessories	65.4	0.2%
Motor vehicle fuels, lubricants, and fluids	371.2	1.1%
Motor vehicle maintenance and repair	176.7	0.5%
Other motor vehicle services	77.6	0.2%
Consumer time usage for maintenance and repair (dollar equivalent)	141.1	0.4%
Vehicle maintenance and repair not done by self	28.2	0.1%
Vehicles	112.9	0.3%
Consumer transport time (B2.3)	**3329.6**	**10.2%**
Consumer time usage (dollar equivalent)	3329.6	10.2%
Travel (for work)	761.9	2.3%
Travel (other)	2567.8	7.8%
Commercial, industrial, and other maintenance, repair, and energy (B1.1)	**170.3**	**0.5%**
Gasoline**	170.3	0.5%
Commercial, industrial, and other purchases (B1.1)	**122.1**	**0.4%**
New motor vehicles	122.1	0.4%
Infrastructure	**76.9**	**0.2%**
Highways and streets	76.9	0.2%
Total - Resources related to discrete manufactured products	4896.8	14.9%
Total - annual resources (GDP and uncompensated labor time)	32771.5	100.0%

* Adjusted to 2014 using the Consumer Price Index for all consumers

** Assumes the ratio of new motor vehicle purchases to Motor vehicle fuels, lubricants, and fluids is the same for consumers as it is for commercial and industrial uses

Source: Bureau of Economic Analysis. (2016) Personal Consumption Expenditures by Function. Table 2.5.5. https://www.bea.gov/itable/index.cfm.
Source: Energy Information Administration. 2009 Residential Energy Consumption Survey. Table CE4.11. https://www.eia.gov/consumption/residential/data/2009/index.php?view=consumption.
Source: Bureau of Labor Statistics. (2016) American Time Use Survey. Table A-1. https://www.bls.gov/tus/a1_2016.pdf.

Table 3: Time Spent Utilizing/Maintaining Transportation Equipment, 2016

Category of time use	Hours per year	Percent of hours awake	Annual time value* ($Billion)
Travel (personal care)	10.95	0.2%	84.65
Travel (eating and drinking)	40.15	0.7%	310.39
Travel (household activities)	18.25	0.3%	141.09
Travel (purchasing goods/services)	105.85	1.9%	818.30
Travel (care of others)	51.10	0.9%	395.04
Travel (for work)	98.55	1.8%	761.87
Travel (for education)	10.95	0.2%	84.65
Travel (for leisure and other)	94.90	1.7%	733.65
Vehicle maintenance and repair not done by self	3.65	0.1%	28.22
Vehicles	14.60	0.3%	112.87
TOTAL	448.95	8.1%	3470.72

* Applying the average hourly compensation for 2014 ($32.305/hour)
Source: Bureau of Labor Statistics. (2016) American Time Use Survey. Table A-1. https://www.bls.gov/tus/a1_2016.pdf.

Figure 3: Environmental Impact and Value Added for the Annualized Life-Cycle Cost of US Passenger Transportation, Percentile

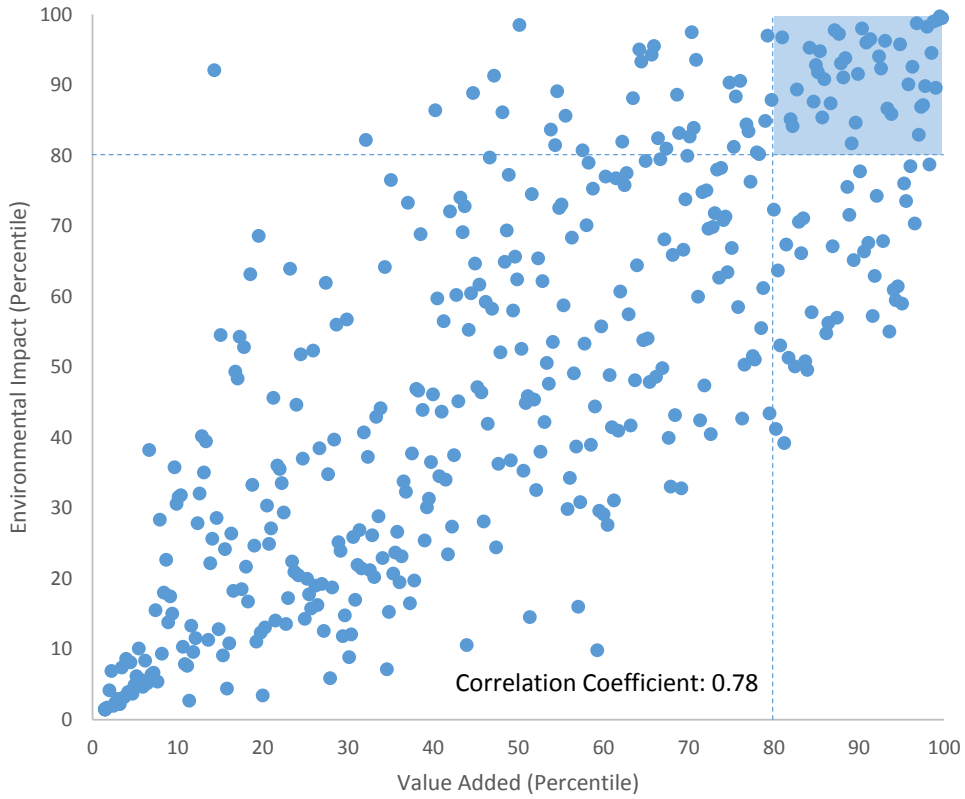


Table 4: Environmental Impact and Value Added for the Annualized Life-Cycle Cost of US Passenger Transportation, Percentile

		Environmental				Value Added			
		Base Case	Mean	Minimum	Maximum	Base Case	Mean	Minimum	Maximum
211000 Oil and gas extraction	**	99.5	99.5	99.5	99.5	99.8	99.8	99.8	99.8
212100 Coal mining		92.9	92.7	91.1	94.1	85.0	84.9	81.0	87.7
2122A0 Iron, gold, silver, and other metal ore mining		96.8	96.8	96.1	97.5	81.0	81.2	76.1	85.2
21311A Other support activities for mining	*	94.1	94.2	93.8	94.6	92.4	92.3	88.4	94.3
221100 Electric power generation, transmission, and distribution		97.8	97.9	97.5	98.3	87.2	87.1	84.2	90.1
230301 Nonresidential maintenance and repair	*	92.6	92.7	91.6	93.6	96.3	95.9	94.3	96.3
233293 Highways and streets		86.9	86.8	85.5	87.9	97.3	97.3	96.8	97.8
327200 Glass and glass product manufacturing	*	98.0	97.9	97.5	98.0	90.4	90.1	86.7	92.9
331110 Iron and steel mills and ferroalloy manufacturing	**	98.8	98.7	98.5	98.8	96.8	96.8	96.6	97.3
331200 Steel product manufacturing from purchased steel		90.9	90.9	89.2	92.1	86.0	85.8	82.3	88.4
33131A Alumina refining and primary aluminum production		91.9	91.8	90.9	93.3	85.2	85.2	81.0	87.9
331411 Primary smelting and refining of copper		85.5	85.5	84.0	86.7	85.7	85.8	81.8	88.4
331419 Primary smelting and refining of nonferrous metal (except copper and aluminum)		86.0	85.9	84.5	86.9	93.8	93.9	91.1	95.3
331490 Nonferrous metal (except copper and aluminum) rolling, drawing, extruding and alloying		89.4	89.2	88.4	90.1	82.8	82.6	77.6	85.7
331510 Ferrous metal foundries		91.1	91.3	90.6	92.1	88.2	88.3	85.0	91.6
331520 Nonferrous metal foundries		93.8	93.8	93.1	94.3	88.4	88.6	85.2	91.9
332310 Plate work and fabricated structural product manufacturing		87.4	87.4	86.2	87.9	86.7	86.7	84.0	89.9
332710 Machine shops		81.8	81.7	80.0	83.5	89.2	89.4	86.0	92.6
332720 Turned product and screw, nut, and bolt manufacturing		86.7	86.6	85.5	87.4	93.3	93.0	89.9	94.8
332800 Coating, engraving, heat treating and allied activities		85.2	85.1	83.7	86.5	82.0	82.0	77.1	85.5
33299B Other fabricated metal manufacturing		84.2	84.3	83.0	85.2	82.3	82.1	77.1	85.5
333618 Other engine equipment manufacturing	*	90.1	90.1	89.4	91.1	95.8	95.7	93.8	96.3
33441A Other electronic component manufacturing		91.6	91.7	90.9	92.6	89.9	89.7	86.2	92.6
336111 Automobile manufacturing	**	99.3	99.3	99.3	99.3	99.3	99.1	98.3	99.3
336310 Motor vehicle gasoline engine and engine parts manufacturing		89.7	89.8	88.9	90.6	99.0	99.0	98.3	99.3
336350 Motor vehicle transmission and power train parts manufacturing		83.0	82.9	81.3	84.5	97.0	97.1	96.8	97.5
336390 Other motor vehicle parts manufacturing		87.2	87.4	86.5	87.9	97.5	97.5	97.0	98.0
316000 Leather and allied product manufacturing	*	96.1	96.1	95.8	97.0	90.9	90.6	86.9	93.3
324110 Petroleum refineries	**	99.8	99.8	99.8	99.8	99.5	99.5	99.5	99.5
325110 Petrochemical manufacturing	*	96.6	96.5	96.1	97.0	91.4	91.0	87.4	93.6
325190 Other basic organic chemical manufacturing		97.3	97.3	96.8	97.5	87.7	87.7	84.7	90.6
325211 Plastics material and resin manufacturing		94.8	95.0	94.8	95.6	85.5	85.6	81.8	88.2
326190 Other plastics product manufacturing	*	92.4	92.7	91.6	93.6	92.6	92.5	88.9	94.8
326210 Tire manufacturing		87.7	87.6	86.7	88.2	84.7	84.7	80.5	87.4
420000 Wholesale trade	*	94.6	94.5	94.1	94.8	98.5	98.6	98.0	99.3
482000 Rail transportation		93.1	92.8	91.4	93.8	87.9	88.0	85.0	91.1
484000 Truck transportation	*	95.8	95.8	95.6	96.1	94.8	94.7	92.6	96.3
485000 Transit and ground passenger transportation	**	98.3	98.3	98.0	98.3	98.0	98.0	97.5	98.3
486000 Pipeline transportation	*	96.3	96.3	96.1	97.0	93.1	92.8	89.4	94.8
550000 Management of companies and enterprises		89.9	90.0	88.9	90.9	97.8	97.8	97.3	98.0
561700 Services to buildings and dwellings		84.7	84.3	80.5	86.5	89.7	89.6	86.2	93.1
562000 Waste management and remediation services		95.3	95.4	95.1	95.6	84.2	84.0	79.8	86.7
811100 Automotive repair and maintenance	**	99.0	99.0	99.0	99.0	98.8	98.7	98.3	99.3

** Both environmental impact and value added are above the 95th percentile

* Both environmental impact and value added are above the 90th percentile, but below the 95th

Facilitation of Smart City and Community Technology Convergence

Martin J. Burns
Smart Grid and Cyber-Physical Systems Program Office
National Institute of Standards and Technology
Gaithersburg, MD United States
martin.burns@nist.gov

Sokwoo Rhee
Smart Grid and Cyber-Physical Systems Program Office
National Institute of Standards and Technology
Gaithersburg, MD United States
sokwoo.rhee@nist.gov

***Abstract*— The National Institute of Standards and Technology (NIST) has led a series of efforts designed to propel consensus toward reusable, standards-based smart city solutions through open collaborations with worldwide participation. This paper describes the novel strategy and methods used in these activities.**

Keywords—Smart City, Architecture, Framework, Composable, Interoperable, Internet of Things, Cyber-Physical Systems

I. INTRODUCTION

Cities and communities of all sizes and types seek to use advanced technologies to make communities safer and more secure, livable, and workable. There is power and value in the propagation of emerging cyber-physical systems (CPS) and Internet of Things (IoT) applications into smart communities. The global smart cities market size is projected to grow from USD 425 Billion in 2017 to USD 1.2 Trillion by 2022 [2].

However, for these solutions to be deployed, some degree of interoperability must be achieved to convince stakeholders that they will not be locked into a single vendor or vendor ecosystem, and, to reduce the costly barriers to integration of new features and capabilities.

For cities and their residents, interoperability is needed to provide for reduced costs, evolvability and extensibility, customization through modularity, expanded range of options and choices, and access for small/rural communities. For innovators and entrepreneurs, interoperability is desirable to enable expanded markets, opportunities for startups and small and medium enterprises (SMEs) including those that produce components but not end-to-end solutions, and platforms for innovation.

Matt Turck in 2016 [1] identified over 1200 organizations directly involved in providing technology to the IoT space which includes smart cities. There are dozens of organizations pursuing IoT and smart cities standards. Additionally, the availability of low-cost kits of powerful computing platforms makes it easy to "invent" an application overnight. [20] As a result, most smart city deployments are built with ecosystems of collaborating service providers that can principally interoperate only with themselves. Convergence among these ecosystems is needed because this diversity is an obstacle to the penetration of these technologies due to lack of replicability, composability and fears of vendor lock-in.

Diversity is a naturally occurring and positive property. However, uniformity is beneficial for lowering cost and achieving economies of scale. For several years, it has been apparent that widespread penetration of specific technologies into smart cities has indeed been hampered by lack of consensus resulting in many applications not progressing beyond pilot stage [3].

IoT, and more generally CPS, pose important measurement and interoperability challenges since smart city applications are inherently cross-sector, multi-technology, and at-scale instances of IoT. Overcoming these obstacles, and achieving the corresponding properties, can be facilitated through interoperability standards. An emerging global smart city technologies market is a growth opportunity for US industry, but conflicting local or regional standards could make it difficult for US companies to compete – especially SMEs.

NIST's Smart Grid and Cyber-Physical Systems Program Office (SGCPS) works to develop and extend the foundations and measurement science for CPS. The SGCPS considers smart cities to be a key opportunity to study CPS and IoT at scale incorporating all the complexities of cross-domain and cross-ownership collaborations between devices, applications, and humans.

In support of this mission and its research interests in Smart Cities and Cyber-Physical Systems, NIST plays a facilitator role in supporting smart city stakeholders and the evolution of IoT technologies deployed in smart cities and communities. This paper describes activities to assist cities and their technology service providers in achieving replicable and scalable smart city deployments. The overall goal is to foster convergence around best practices for smart cities by encouraging the collaboration between stakeholders for the common good and economies of scale.

How can a small part of a small agency have a positive impact in a large economic sector like smart cities? Overall, NIST has employed a unique stakeholder engagement strategy. NIST's strategy was to organize teams of cities/vendors/government/academics, giving voice in particular to cities to explain their needs and have partners

Rhee, Sokwoo; Burns, Martin.
"Facilitation of Smart City and Community Technology Convergence."
Paper presented at Third International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership with Global City Teams Challenge (GCTC), Porto, Portugal. April 10, 2018 - April 13, 2018.

respond to those needs. The teams were focused enough to make tangible progress on an identified opportunity.

This NIST strategy was that of growing collaborative concentric circles of stakeholders allowing them to absorb designs and concepts from one and other. This was also consistent with our technical strategy, which involved using systems engineering principles (NIST's CPS Framework [9]) to work bottom-up from device-level performance characterization to system performance to connected systems to infrastructures (local scale) to extreme-scale complex connected infrastructures.

Specifically, the problems to be addressed were: (1) costly, constrained custom solutions and associated market fragmentation and stranded investments [4]; (2) lack of interoperability; and (3) disjointed standards efforts.

The SGCPS approach to addressing each of these three problems are:

(1) Create forces for convergence:

   a. Connect cities/communities to one another to work together,

   b. Promote public-private partnerships that join industry and academia with city partners, and

   c. Nurture, identify, and replicate success.

(2) Identify emerging points of convergence, or Pivotal Points of Interoperability (PPI), in existing deployments and architectures.

(3) Use the CPS Framework [9] as a 'Rosetta Stone' to map the various standards efforts to one another, identify standards gaps, and facilitate prioritization of standards efforts.

Success can drive stakeholder convergence around best practices, interoperability, composability for smart city applications. This in turn speeds the penetration of these IoT and CPS applications so that the social and economic benefits can be realized.

The balance of this paper describes efforts by NIST's SGCPS to help facilitate voluntary convergence of applications of smart features in cities and municipalities.

II. Create Forces for Convergence

In order to encourage cooperation and coordination among stakeholders – communities, businesses, academic institutions, and non-profit organizations, SGCPS has undertaken a sequence of collaboration projects. These projects address the concerns of smart city propagation from two directions:

(1) A market-driven component that provides a place for smart cities and their vendors to collaborate and exchange lessons learned and best practices from their experiences; and

(2) A technology-driven component through analysis of deployment architectures and requirements analysis of smart city features.

Over the course of several years, these efforts have been successful at increasing the ability of smart city applications to be deployed and replicated, without picking winners or losers or making value judgements about the participants.

NISTs smart city convergence efforts began with the Smart America Challenge Workshop, held with the support of the Office of Science and Technology Policy (OSTP) Presidential Innovation Fellows program at the end of 2013. [5][6] This effort broadened in subsequent years to become the NIST Global City Teams Challenge (GCTC) [11] and the Internet of Things Enabled Smart City Framework (IES-City Framework). [19]

What follows is a description of the components of these activities that were brought to bear on the smart cities challenge.

A. Partnering with Stakeholders

A key element of NIST's approach has been strategic engagement with multiple classes of stakeholders:

Partner with agencies – Several agencies in the US federal government have smart city activities including OSTP, Department of Transportation (DOT), Department of Homeland Security (DHS), National Science Foundation (NSF), Department of Energy (DOE), National Telecommunications and Information Administration (NTIA), and International Trade Administration (ITA).

Partner with cities and communities – Through Global City Teams Challenge activities, smart city proponents and pioneers are provided a forum in which to discuss and compare their efforts to mutual benefit and leverage each other's investments and knowledge. GCTC teams with multiple cities also enable development of common sets of requirements to support development of more comprehensive and scalable solutions applicable to a broad set of cities and communities, thus increasing potential market size.

Partner with technology providers – Through the challenge activities, technology providers gain an opportunity to showcase their solutions and learn from each other.

Incubation of projects – Projects that begin as small pilots and even academic research can progress through stages of iteration, maturing, and gaining acceptance.

Kickoff / match making – A forum is created for potential ecosystem collaborators to meet and join to address potential applications together.

Expos – Expositions provide all collaborators the opportunity to showcase and learn from each other's achievements.

B. Nurturing of participants

Through the course of engagements in the various collaborative activities, active engagement with the participants and teams is achieved through teleconferences, email, in-person presentations and small workshops. Cities and communities in search of best practices with the goal to address common issues are encouraged to collaborate to deploy shared solutions. Technology providers and researchers establish project teams through partnership with cities and communities to demonstrate the value of their capabilities. Once successful examples are identified, technology providers are encouraged to work with additional cities and communities to replicate their success.

Additionally, several grant award opportunities from NIST and NSF were made available for the teams to jumpstart building partnerships and accelerate research and development.

Rhee, Sokwoo; Burns, Martin.
"Facilitation of Smart City and Community Technology Convergence."
Paper presented at Third International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership with Global City Teams Challenge (GCTC), Porto, Portugal. April 10, 2018 - April 13, 2018.

C. Timeline

Over the course of three years, the market-driven component sequentially progressed with broader and broader degrees of convergence and interoperation of smart city applications. Figure 1: Successive Convergence in GCTC illustrates these achievements.

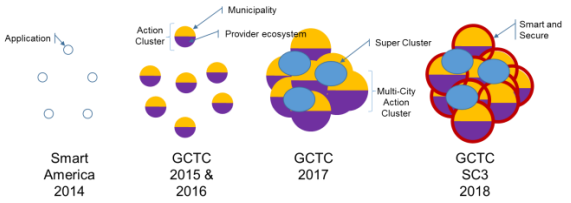


Figure 1: Successive Convergence in GCTC

1) Smart America

The Smart America Challenge (SmartAmerica) attempted to address the issue of fragmentation in IoT and CPS technologies and applications. Specifically, SmartAmerica sought to bring together organizations with IoT and CPS technologies, programs, and testbeds with the goal of demonstrating the potential of multi-domain collaboration of IoT and CPS to create tangible economic benefits, create jobs, save lives, and improve the overall quality of life.

Smart America was able to assemble over 100 commercial and academic organizations and government agencies in this collaboration. The program lasted a year culminating in a showcase event in 2014 with 24 cross-disciplinary teams, each of which was composed of multiple organizations. Participating teams were asked to develop cross-domain applications such as a "Crash-to-Care" scenario, where the victims of massive traffic accidents could be efficiently triaged and transported to appropriate medical facilities in a seamless manner. Through the process, SmartAmerica tried to identify cross-cutting themes shared by multiple IoT/CPS applications in different domains. SmartAmerica was also able to demonstrate a number of cross-domain applications with the potential for IoT and CPS to create tangible economic benefits or to save lives. For example, the Smart Emergency Response Team (SERS) [7] demonstrated that a combination of robots, sensor-equipped dogs, drones, and a command center could effectively collaborate to deal with several emergency response scenarios. The Closed-loop Healthcare team [8] demonstrated the importance of interoperability between in-hospital and in-home health monitoring systems, in providing better healthcare experiences while saving costs. Through such examples identified and incubated in its process, SmartAmerica demonstrated that it was not only possible, but also necessary, for different applications to collaborate to unleash the true potential of IoT and CPS technologies.

2) Global City Teams Challenge

Based on the success of Smart America, and recognizing that IoT and CPS applications were being deployed throughout the cities and communities around the world, the Global City Team's Challenge (GCTC) was created. [11]

GCTC was devised to provide a forum for demonstrations of smart city applications with tangible benefits to community residents. The collaborators in these demonstrations were provided with facilitation resources, encouragement, and a means to demonstrate their results in an annual conference.

A couple of important use cases are instructive. Imagine a new entrant into the smart city application space. Any prospective customer might ask a technology provider – "where is your technology deployed?" This sets up a "catch 22 [1] " situation. However, through GCTC this technology provider can be part of a team deploying a pilot application exposed at a GCTC exposition. And thus, the catch is resolved. Another case is of a team with a great idea and a willing municipality to try the idea. However, the absence of a critical component or skill prevents moving forward. At a GCTC "match making" event, the gap can be filled and all can pursue the opportunity together. Finally, even mature and well-funded participating enterprises can benefit from the availability of recognition in a forum where such achievements can be viewed by both existing and prospective customers.

Through the GCTC process, cities and communities can help each other find and replicate proven solutions which might have already produced successful results in other municipalities, at a lower cost than developing a similar solution from scratch. Technology providers can replicate successful solutions to a larger number of cities and communities and benefit from the economies of scale.

The design of GCTC includes the facilitated formation of "action clusters" which are teams of collaborators working voluntarily on one or more smart city deployments. Each action cluster consists of a host city or cities and the team of technology providers and researchers that together will realize the deployment.

Initial action clusters meet at an annual kick-off workshop where discussion and matchmaking occurs. During the course of the challenge, these groups make progress, are enhanced by interactions and addition of new team members, and new action clusters join by associating their efforts with GCTC.

At the end of the challenge year, the participants share in a showcase expo where their achievements can be viewed and otherwise celebrated.

These features enabled GCTC to attract hundreds of projects and a willingness to converge within this venue. The 2015 instance of GCTC featured 64 action clusters including over 50 cities from around the world at its expo. The 2016 version had 100 action clusters and 110 cities.

Based on the success of this GCTC model in its first year, NIST sought to increase the degree to which these deployments could be made more replicable. The experience with the action cluster teams of previous challenges revealed several recurring themes and classes of applications. NIST recognized the opportunity for further convergence by creating the notion of a "SuperCluster."

SuperClusters represent groups of action clusters around a common theme. For example, a transportation supercluster was formed to align the efforts of several smart city transportation related projects. To date super clusters have been formed as follows: Transportation, Public Safety, Utility (Energy, Water, Waste Management,) Data Platform/Dashboard, Public WiFi/Wireless, Data Governance/Exchange, Agriculture/Rural, and Education.

[1] Catch 22 refers to the notion expressed in a satirical novel by Joseph Heller that considers the enigma of a goal that can't be achieved because it requires the prior achievement of said goal.

Each SuperCluster was encouraged to bring together multiple cities and applications which could be designed and replicated across multiple locales. These designs have been captured in "blueprints" which are published documents describing the common requirements and concepts behind the projects. Although every city and community is different to an extent, there are a number of common issues shared by groups of municipalities. SuperCluster blueprints document the technologies that can address common issues and help the cities and communities jumpstart planning and deployment of replicable and successful best practices without going through the painful and complicated process that other cities may have already gone through. During the 2017 round, GCTC SuperClusters have published 5 blueprints [12]. The action clusters nurtured in GCTC are actively replicating their solutions in multiple cities. Examples include Array of Things [13] being deployed in Chicago, Illinois and Portland, Oregon. The kiosk technology initially deployed in New York City is now also adopted by Bexar County, Texas [14]. The concept of the SuperCluster and its blueprints are referenced and modeled by the Virginia Smart Communities Working Group [15].

GCTC was able to attract several government and non-profit grant making organizations that helped fund several of the cluster activities. Included were NSF Early-concept Grants for Exploratory Research (EAGER), and NIST funded Replicable Smart City Technology (RSCT) grants. The goal of NIST's RSCT grant program was to identify and nurture the technologies with the best potential of replication in multiple cities and communities. RSCT grantees were encouraged to work with more than one municipality to share their solution. One of the NIST grantees was the StormSense team composed of seven cities and counties in Southern Virginia. The team has developed an inundation forecasting technology that can estimate water level rise during flooding events. Since flooding typically affects multiple municipalities, it made sense for the cities to jointly develop and deploy an interoperable solution that can be shared together. Although a variety of wireless technologies were adopted by different jurisdictions (WiFi, LoRa, cellular), collected sensor data were commonly brought into a shared cloud platform and injected into the same analytics model, which produced a comprehensive forecast covering a broader region than a single municipality [16]. Another NIST grantee working with Montgomery County, Maryland, has developed a data sharing and exchange platform that could easily connect multiple applications covering different domains such as transportation, agriculture and healthcare [17]. Multiple cities and communities including Washington DC are considering adoption of the platform. One of the action clusters in the 2015 and 2016 rounds, the Smart Mobile Operation OSU Transportation Hub (SMOOTH) project from the Ohio State University that tested a network of on-demand automated vehicles, became a core component of the proposal from the City of Columbus, Ohio, winning the $40 Million DOT Smart City Challenge in 2016 [18].

3) GCTC-SC3

SmartAmerica and GCTC have successfully nurtured and documented replicable smart city deployments over four years. The growing number of innovations, however, cannot be practically adopted at scale without serious considerations of security and privacy. Cities and communities are aware of the importance of planning for cybersecurity, privacy, and trustworthiness risks in their IoT and CPS deployments, but many of them lack a clear vision and the expertise to address risks in a systematic manner. Industry stakeholders are eager to address the issues in their products and solutions as well, but many of them struggle to find a clear business and engagement model with city and community customers.

In the 2018 round of GCTC, NIST has partnered with the US Department of Homeland Security (DHS) Science and Technology Directorate (S&T) to tackle these issues as a first order concern. DHS S&T has a strong track record of working with cybersecurity and privacy professionals and possesses abundant internal and external expertise on the relevant issues. Building on GCTC, NIST SGCPS and DHS S&T Cybersecurity Division decided to jointly run a 12-14-month program for teams of cities and innovators to demonstrate value and return on investment for designed-in trustworthiness for smart city deployments. This new program has been named the Smart and Secure Cities and Communities Challenge (SC3). In GCTC-SC3, action clusters are required to describe and demonstrate their considerations of security and privacy as well as the replicability and practical impacts of their solutions. It is the goal of the program to facilitate introduction of best practices for good security and privacy measures into the domain-specific SuperCluster activities and update their blueprints to include strong flavors of security and privacy. These blueprints can be used by cities and communities in adopting replicable, secure, trustworthy, and privacy-enhancing IoT and CPS solutions.

Throughout the 4 rounds of SmartAmerica and GCTC, NIST has built a strong community of smart city and IoT stakeholders who are willing to collaborate and deploy solutions in partnerships. In fact, over 180 action clusters composed of more than 400 technology providers and 160 municipal governments have participated in SmartAmerica and GCTC. At the time of this writing GCTC-SC3 is underway.

III. Identify Emerging Points of Convergence

While GCTC can be considered a market-driven effort at driving smart city convergence, a new activity was begun in 2016 called the Internet of Things Enabled Smart City Framework (IES-City Framework). The goal of this effort was to provide an impetus towards technical convergence. The IES-City Framework was released at the beginning of 2018 as a draft for review and anticipates a formal release around summer of that year.

A collaboration sponsored by eight national and international partners, IES-City Framework convened technology providers and researchers to define a taxonomy and methodology for comparing smart city applications and technical components.

IES-City established two key principles to help in this analysis: Pivotal Points of Interoperability (PPI) and Zones of Concern (ZofC). Additionally, a tool was created to enable the rapid review of smart city applications for their breadth and functional requirements, the readiness of a city or municipality infrastructure to mount or absorb applications, and the benefits to the city or municipality from these applications.

Discovering PPI

Pivotal Points of Interoperability is a powerful concept that recognizes that when you standardize everything, innovation can be frozen out; if you standardize nothing, interoperability will not be achievable. There is a range of optimal convergence in between these two extremes. There are many architectures and technology components in application of smart cities as

previously established. Yet, on careful inspection, it can be seen that many such technology components were built with common building blocks: not through coordination but through similar independent technical choices in the absence of coordination. An easy example is the adoption of Internet Protocol [21] for the identification of endpoints in a communications network and a means to routing messages to them. Virtually all smart city and IoT applications rely on this PPI.

IES-City conceived that there are numerous such common decisions that were made independently by smart city decision makers in the design of smart city applications.

In order to expose these "consensus in place" choices, an analytical approach was devised using the *aspects* and *concerns* from the NIST CPS Framework [9]. The concept is that for any given part of a system or system of systems there is a set of concerns that are being addressed: privacy of data is a good example. The CPS Framework derived nine groupings of concerns termed *aspects* including functional, business, human, trustworthiness, timing, data, boundaries, composition, and lifecycle *aspects*. Within those *aspects* are over 100 individual subsidiary concerns. For example, the trustworthiness *aspect* consists of several concerns at the next hierarchical level: security, privacy, safety, reliability, and resilience. Note how this cluster of concerns encapsulates the family of concerns about the avoidance of harm in the design and deployment of CPS.

Figure 2: Revealing PPI

Using this concept, proponents of architectures and technology suites prominent in smart cities were invited to analyze their designs via a table where the rows were CPS Framework concerns (see Figure 2 Revealing PPI). While this approach does not intend to collate all the detail of their designs, it does serve the purpose of exposing the substantial choices made in technologies to address the concerns. By offering the ability to present these analyses side by side, it is anticipated that PPI can be revealed, i.e., common choices made in addressing a common concern.

Zones of Concern (ZofC)

As one reviews the concerns from the CPS Framework and the analyses provided by the proponents of technology suites, sets of services that address related sets of concerns emerges. These so-called zones of concern (ZofC) can result in service offerings that can simplify the distribution of responsibilities among teams working to deploy smart city applications – this is a form of convergence itself. For example, there are typically three kinds of collaborators in the deployment of smart city applications – the device vendor that makes IoT sensors and actuators and such, the application designer that composes information and provides the visible benefit to the citizens, and the infrastructure provider that maintains the glue that allows applications to be device agnostic and the device vendors to be application agnostic.

When a reviewer surveys the many architectures being proposed and deployed for IoT and smart cities, these distinct boundaries of responsibility can be seen in the architectural diagrams. Typical is the illustration of a "northbound interface" where applications including human presentation and analytics connect, and, a "southbound interface" where devices connect. Finally, the "glue" is an infrastructure where assembled sets of services addressing concerns can be made available at the appropriate northbound and southbound interfaces.

Application Framework Tool

Finally, to simplify initial research into smart city applications by less-technical stakeholders, an application framework tool was constructed around a set of categories and subcategories of smart city applications observed in deployments in GCTC and elsewhere around the globe. The categories and subcategories represent the breadth of known smart city applications.

These categories were analyzed in each of three dimensions:

Breadth, and functional and ICT requirements: For this dimension, each subcategory was analyzed against the CPS Framework *aspects* and concerns for high level requirements for their realization.

Readiness required by city infrastructures and citizenry to enable the mounting of these applications: rather than a complete set of metrics or maturity characteristics, this subset of key indicators was reviewed against each subcategory to determine general support for the potential deployment.

Benefits to the citizenry and city from the deployment of these applications: each subcategory was analyzed for the public sector, private sector, and citizenry benefits afforded. And within each benefit grouping economic, environmental, and social benefits were assessed.

Together, this analysis tool facilitates early evaluation of smart city technologies by smart city stakeholders allowing for optimized planning and specification of the evolution of their locales.

IV. The NIST CPS Framework as a Rosetta Stone

In communicating about an application, a terminology is helpful. If the terminology is specific to the application, the interpreter must first master the terminology before understanding what the terminology is used to describe. Unless this terminology is more widely used, the learning curve for understanding in each instance is a barrier to understanding. NIST recognized this need and addressed the design of such a Rosetta Stone of terminology for describing CPS and IoT. The NIST CPS Framework provides this common set of terminology.

CPS is an inherently complex topic because it is cross-cutting over enterprises, function, and technologies. For this reason, NIST convened a CPS public working group which produced the material allowing a NIST Special Publication of the CPS Framework in 2017 [9].

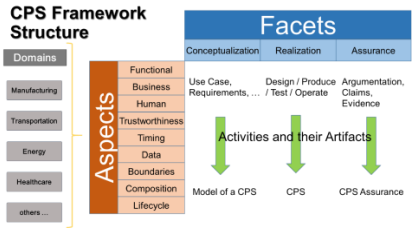


Figure 3: CPS Framework

As shown in Figure 3: CPS Framework, the CPS Framework identifies two axes of definition: *aspects* and *facets*.

Aspects allow for the categorization of concerns for which a CPS/IoT/Smart City application needs to address. It is asserted that all possible concerns that drive services are potentially in mutual support or conflict and therefore need to be treated uniformly. The *aspects* are further decomposed into concerns and sub-concerns producing a "concern tree." For example, the Trustworthiness *aspect* is comprised of concerns about security, privacy, safety, reliability and resilience. Security is broken down further into physical and cyber. And this pattern then repeats. *Facets* represent modes of thinking about CPS. They categorize activities performed over the lifecycle of the CPS. For example, the conceptualization *facet* includes activities such as business and use case development and requirements analysis. The result is a model of a CPS. The realization *facet* deals with activities that make up the design, implementation and deployment of the CPS. As such its result is the CPS itself. Finally, the assurance *facet* is about activities that result in an assurance case that the CPS performs as desired. The result of these activities is an assured CPS.

By using these simple concepts, the pertinent characteristics of any CPS and therefore smart community application can be discussed in a way that it is comparable with any other such description.

The result is that the many disjoint efforts to describe diverse smart city, IoT, and CPS applications can use this Rosetta Stone to reduce the barrier to understanding of what is described. The IES-City Framework made substantive use of this concept in its technical analyses.

V. Conclusion

Taken together, the activities facilitated by NIST and described in this paper provide a direction toward convergence in smart cities technologies. These results lower the cost and complexity of deploying smart city applications and importantly provide a means of growing them beyond their initial scope to add additional features as they become feasible and available.

NIST's approach of convening stakeholder groups allows this convergence to occur naturally and in open non-discriminatory forums to the benefit of all participants.

The use of the CPS Framework and common application documentation from IES-City Framework and GCTC Superclusters further inspire and inform NIST research into CPS and IoT.

For additional information on GCTC activities see https://pages.nist.gov/GCTC.

For additional information on IES-City Framework activities see https://pages.nist.gov/smartcitiesarchitecture/.

Acknowledgment

The authors wish to acknowledge the many colleagues at NIST, government agencies, IES-City Framework partners, and individual organizations who collaborated with us on CPS Framework, SmartAmerica, GCTC, and IES-City Framework.

References

[1] Turck, M (2016), Blog, *Internet of Things: Are We There Yet? (The 2016 IoT Landscape)*, retrieved on 11/17/2017 from http://mattturck.com/2016-iot-landscape/

[2] Retrieved from https://www.marketsandmarkets.com/Market-Reports/smart-cities-market-542.html

[3] Van Winden, W. (2016). *Smart city pilot projects, scaling up of fading out? Experiences from Amsterdam*. Regional Studies Association Annual Conference in Austria, Graz, 3rd – 6th April, 2016

[4] Rhee, S., " Catalyzing the Internet of Things and Smart Cities: Global City Teams Challenge," 2016 1st International Workshop on Science of Smart City Operations and Platforms Engineering in partnership with Global City Teams Challenge, Pages 1-4, Vienna, Austria, April 11, 2016, IEEE Explore Digital Library

[5] Retrieved from https://obamawhitehouse.archives.gov/blog/2014/06/10/smartamerica-challenge-harnessing-power-internet-things

[6] Retrieved from http://smartamerica.org/

[7] Retrieved from http://smartamerica.org/teams/smart-emergency-response-system-sers/

[8] Retrieved from http://smartamerica.org/teams/closed-loop-healthcare/

[9] E. Griffor, C. Greer, D. Wollman and M. Burns, "Framework for cyber-physical systems: volume 1, overview," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication. NIST-SP-1500-201, June 26, 2017. doi: 10.6028/NIST.SP.1500-201

[10] ISO/IEC JTC 1 WG11 Smart Cities, http://www.iec.ch/dyn/www/f?p=103:14:0::::FSP_ORG_ID,FSP_LANG_ID:12973,25

[11] Retrieved from https://pages.nist.gov/GCTC/

[12] Retrieved from https://pages.nist.gov/GCTC/super-clusters/

[13] Retrieved from http://arrayofthings.github.io/

[14] Retrieved from http://civiqsmartscapes.com/press/civiq-smartscapes-pens-landmark-deal-with-bexar-county-to-transform-downtown-san-antonio-into-connected-smart-community

[15] Retrieved from https://governor.virginia.gov/newsroom/newsarticle?articleId=21810

[16] Loftis, J. D., Wang, H., Forrest, D., Nguyen, C., Rhee, S., "Emerging Flood Model Validation Frameworks for Street-Level Inundation Modeling with StormSense," Proceedings of the 2nd International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE), Pages 13-18, Pittsburg, April 18-21, 2017, ACM Digital Library

[17] Nelson, A., Toth, G., Hoffman, D., Nguyen, C., Rhee, S., "Towards a foundation for a collaborative replicable smart cities IoT architecture," Proceedings of the 2nd International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE), Pages 63-68, Pittsburg, April 18-21, 2017, ACM Digital Library

[18] Retrieved from https://www.transportation.gov/sites/dot.gov/files/docs/Columbus%20OH%20Vision%20Narrative.pdf

[19] Retrieved from https://pages.nist.gov/smartcitiesarchitecture/

[20] Postscapes. (2018). *IoT Hardware Guide*. Retrieved from https://www.postscapes.com/internet-of-things-hardware/

[21] IETF (1981) *Internet Protocol* (RFC 791). Retrieved from https://tools.ietf.org/html/rfc791

Smart and Secure Cities and Communities

Scott Tousley
Cybersecurity Division
U.S. Department of Homeland Security, Science and Technology Directorate
Washington, DC United States
scott.tousley@hq.dhs.gov

Sokwoo Rhee
Smart Grid and Cyber-Physical Systems Program Office
National Institute of Standards and Technology
Gaithersburg, MD United States
sokwoo.rhee@nist.gov

***Abstract*— Cities and communities around the world are increasingly deploying advanced technologies such as Internet of Things and Cyber-Physical Systems for improved efficiency, convenience, safety, and quality of life. The Global City Teams Challenge (GCTC) program, led by the National Institute of Standards and Technology, has successfully nurtured hundreds of projects led by municipal governments and technology providers with the goal of helping build smart infrastructure through advanced technologies. However, security and privacy for smart city solutions have not been generally considered a top priority by the smart city community. To address this issue, the U.S. Department of Homeland Security's Science and Technology Directorate has joined forces with NIST in the 2018 round of GCTC to encourage smart cities and communities to embrace security and privacy as a primary concern.**

Keywords— Smart City, Smart Community, Security, Privacy, Internet of Things, Cyber-Physical Systems, Intelligent Infrastructure, Global City Teams Challenge

I. INTRODUCTION

Smart cities use advanced technologies such as Internet of Things (IoT) and Cyber-Physical Systems (CPS) to improve the quality of life for their residents. [1] Although cities and communities around the world are increasingly adopting advanced technologies for improved efficiency, convenience, safety, and security, the smart city market still suffers from the issues of fragmentation and heavy reliance on custom-tailored solutions and one-off projects that lack a sustainable operational model beyond their initial deployment. To address this issue and promote identification and replication of best practices with measurable benefits in cities and communities, the National Institute of Standards and Technology (NIST) launched the Global City Teams Challenge program (GCTC) in 2014. [2] In partnership with other U.S. federal agencies, the program has achieved substantial success benefiting the American public and international participants, through a combination of a strong collaborative approach, and sustained interest by cities and communities of all sizes throughout the nation. Beginning with the first round that culminated on June 2015, the program has supported and sparked hundreds of "smart cities" projects involving a wide range of businesses, academia, government and civic organizations, [3] most of which are operating today. However, the many different GCTC efforts to date have not yet generated much consideration of cybersecurity and privacy, an increasingly significant gap in the program. So, for 2018, NIST and the U.S. Department of Homeland Security, Science and Technology Directorate (DHS S&T) have started a working partnership to greatly expand consideration of cybersecurity and data privacy in the GCTC program.

This partnership is very important to the GCTC program and "smart cities" effort going forward, because the challenges of cybersecurity and privacy continue to grow steadily. These challenges are growing more quickly than our ability to manage and overcome them, especially from the perspective of typical small and medium sized organizations throughout the nation, economy and society. Threat and risk "asymmetries" will remain significant – the character of smart cities, communities and systems is that they will be susceptible to attack and disruption from a variety of active threats based throughout the world. These threat actors will retain the ability and initiative to cause problems when and where they choose. The GCTC- Smart and Secure Cities and Communities Challenge (SC3) program represents the ultimate use case for cybersecurity and data privacy capabilities that are being developed through both the private and public sector's Research, Development, Test, and Evaluation (RDT&E) efforts. These capabilities span areas such as software security and assurance, test and evaluation, CPS, data privacy, critical infrastructure security and resilience, network systems security, identity management, open source technologies, information sharing and analysis, incident response/exercises/red teaming, law enforcement forensics capabilities, cybersecurity risk frameworks, RDT&E collaboration, cybersecurity economics and insurance, and technical and organizational education. NIST and DHS S&T expect that the GCTC-SC3 program will address a number of these topics through the various Supercluster and Action Cluster efforts.

II. SMART CITIES' CYBERSECURITY AND DATA PRIVACY CHALLENGES - COMPLICATED AND COMPLEX

The nation's cities and communities are active laboratories for modern technology applications and deployment – critical infrastructure systems, CPS, various IoT elements, and smart cities. These cities and communities are increasingly both complicated and complex, where the key difference is the degree of interrelated factors and unknowns, and the potential

Rhee, Sokwoo; Tousley, Scott.
"Smart and Secure Cities and Communities."
Paper presented at Third International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership with Global City Teams Challenge (GCTC), Porto, Portugal. April 10, 2018 - April 13, 2018.

for emergent properties to become a significant part of the outcomes in these cities and communities. With the development of smart cities' capabilities, cities and communities throughout the nation will be moving from complicated to complex social/technological systems behavior.

But while most government organizations (at all levels) expect to continue managing their complicated areas with defined systems, processes and hierarchies, it is a desire for emergent, and beneficial economic outcomes that is driving interest in complex smart cities activities throughout the nation. This could be dismissed as an exercise in semantics, except for one thing: When facing a problem, managers tend to automatically default to complicated thinking. Instead, they should be "consciously managing complexity….if you manage complex things as if they are merely complicated, you're likely to be setting your (city/community) up for failure." [4] Consciously managing complexity in a business (or government) context is broadly a function of four different strategies or tactics. They are: (1) recognize which type of system you are dealing with; (2) think "manage, not solve"; (3) employ a "try, learn, and adapt" operating strategy; and finally, and perhaps most importantly, (4) develop a complexity mindset. These are important elements of how to approach smart cities' cybersecurity and data privacy challenges.

There are other important elements of this complicated/complex question of smart cities. The nation's most critical infrastructure and smart cities elements are substantially decentralized in their design, operation and evolution, although there are some elements of reasonable consistency (transportation systems, power grid elements, financial transactions, etc.). Investments in these elements across the country, in both the public and private sectors, are different. These differences and the bottom-up character of smart cities' activities and projects also point to the likelihood of emergent city and community properties rather than just simpler architected outcomes. This is a significant part of the smart cities' cybersecurity and data privacy challenge – rather than a complicated enterprise/risk framework problem, we are engaging a series of complex emergent city and community cybersecurity and data privacy challenges. This points to the need for a complexity mindset, a manage-rather-than-solve approach, and try/learn/adapt strategies.

III. Safety, Security, Privacy and Quality

Another part of the SC3's complexity challenge is the combination of safety, security, privacy and data/information systems activity inherent to smart cities and communities. IoT and CPS, which most smart city projects are based on to some degree, are hybrid systems of physical and logical elements. Due to the interdependency of logical and physical domains, the impact of security and privacy issues on the logical system inherently affect the behavior of the physical system, and vice versa. These characteristics of IoT and CPS create unique challenges and complexities in the development and deployment of smart city solutions. IoT and CPS deployments must take into account the possibility of unintended consequences that may expand beyond the conventional understanding of cybersecurity and privacy. For example, when a vehicle's cybersecurity is compromised, it is no longer a problem confined in the logical world, as the vehicle may be exploited to cause physical damage. In addition, having cybersecurity capability in a device (e.g. vehicle) does not necessarily mean the complete system (e.g. city's overall transportation system) is secure. The issues of safety, security, privacy, and quality must be holistically analyzed and mitigated in terms of the overall concept of the trustworthiness of the system. [5]

Successful smart city deployments require successful safety systems design and operation, as well as in-depth considerations of security and privacy, typically in combination with the treatment of a range of human behavioral expectations in these same designs. The increasing interest in the cybersecurity of smart city designs and operations can leverage the coupling between complex system safety and security, where these two topics are partly overlapped in the typical elements needed in both areas.

Typical discussions of data security orient on confidentiality, integrity and availability of data, versus the appropriate use of data. These connections are increasingly dynamic and fluid in an increasingly data-oriented world where any single individual or organization is necessarily touching multiple sources of data and information elements – which is the nature of our smart cities and communities environments. Smart city stakeholders are increasingly paying attention to the challenge of building a governance and exchange model for IoT data and a plan that governments and technology providers can successfully customize and deploy. It is virtually impossible to exchange and share data in a flexible and trustworthy manner without a consensus-based framework or guideline agreed upon by participating entities. The management and governance of data and information systems will drive all elements of security and privacy outcomes – all happening at large scale and accelerating speed. And all requiring a significant and improving degree of quality.

A significant degree of our national cybersecurity challenge today is from the inconsistent quality of our information technology systems, as designed, deployed and used. Thinking about CPS, IoT, and smart cities, it should be noted that new capabilities are being installed faster than can be recognized and caught up to the large "technical debt" of the systems and capabilities already in place. This is a truly iceberg problem where only a little bit of the problem is visible to us today. The historical disincentives for high-quality information technology systems remain very significant – speed and capability and new features/new users are prized above all, and we continue to draw software, hardware and systems/elements from the existing resource pool that is of limited quality.

The quality definition in the context of this discussion is intended as sustained "fitness for purpose;" the degree to which the systems and capabilities are reliable, maintainable

and flexible, and has the sustained ability to perform satisfactorily in service and is suitable for its intended smart cities' purposes. Quality should be defined as a reasonably strong degree of complex systems of security, safety, and data privacy, and adaptability to both threats and learning. This means catastrophic and major security breakdowns should be avoided, and security risk should be actively managed so that it remains a lesser business problem rather than major corporate risk/threat.

The usual notion of quality is assessed relative to defects in design, i.e., failure of requirements in the design. The important ways to achieve quality are broad collaboration and learning systems, more of an open "no-fault" learning approach than a closed approach where successful security elements are treated as proprietary competitive advantage. It is necessary to learn and adapt as quickly or faster than the threat organizations that are the basis of the problem. Increasing cybersecurity quality in enterprise and broad smart cities environments remains one of the key challenges facing leadership and management at all levels.

One of the key tools in how NIST has approached the challenge of raising cybersecurity quality throughout the nation is through development, deployment and follow-through of the cybersecurity risk framework. [6] Cities and communities can leverage various available resources such as the NIST cybersecurity framework and the cyber-physical systems framework [5] to start engaging in this question, and develop enhanced strategies that would satisfactorily address unique needs in deploying smart city solutions. As GCTC-SC3 looks to understand and drive the smart cities/cybersecurity intersection, many cities and communities may need simpler tools and heuristics. An informal analysis framework could prove crucial in this early effort, such as:

- Understand: Threat environment; Key systems; Connections and interactions
- Evaluate: Systems analysis; Systems testing; Risk assessment
- Adjust: Prioritization; Risk management; Continuous analysis and learning

IV. Smart and Secure Cities and Communities Approach

Cities and communities are generally aware of the cybersecurity, privacy, and other trustworthiness risks in their smart city deployment, but not many of them have a clear vision and the expertise needed to address them.

Industry stakeholders driving smart city projects are eager to address cybersecurity, privacy, and other trustworthiness issues as well, but struggling to find a clear business and engagement model. The GCTC-SC3 program, co-hosted by NIST and DHS S&T, is designed to tackle these issues as the first order concern. Building on the previous rounds of GCTC, NIST and DHS S&T will run a 12-14 month program for teams of cities and innovators to demonstrate value and return on investment for designed-in trustworthiness for smart city deployment, as is illustrated in Figure 1.

To support the GCTC effort, NIST is leveraging a pair of organizational constructs, "Action clusters" and "Superclusters" as shown in Figure 2. Action clusters are Smart Cities projects oriented around particular cities, communities and users/applications with a defined goal. In an Action Cluster, committed cities/communities and partners jointly tackle shared issues, develop and deploy shared solutions to leverage each other's investment and create economies of scale. On the other hand, a SuperCluster is an alliance of action clusters in the same domain such as wireless communications, data governance, transportation, and utilities; and may be joined by individual entities that do not belong to a specific action cluster. A typical SuperCluster is a multi-city, multi-stakeholder collaboration organized around common project objectives and shared solutions. For example, a smart parking project and a traffic congestion management project could be two independent action clusters, but both of them could be included in a transportation SuperCluster. Each SuperCluster produces a "blueprint" to be used by cities and communities around the world as the foundation for building their own smart city strategies. [7] The main goal of the blueprint is to help the cities and communities to jumpstart planning and deployment of replicable and successful best practices without going through the painful and complicated process that other cities may have already gone through.

Figure 1: NIST and DHS S&T Partnership

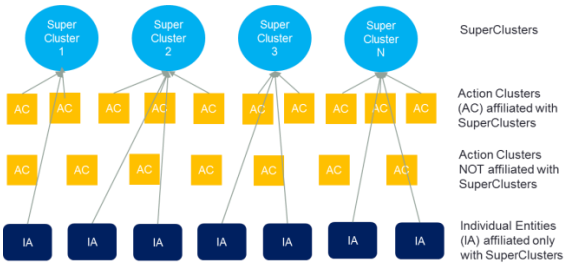


Figure 2 GCTC Organizational Constructs

In developing GCTC-SC3, NIST and DHS S&T agreed that its goal would be best achieved by encouraging the active collaboration of a range of security and data privacy capabilities across all of the Superclusters and Action clusters,

Rhee, Sokwoo; Tousley, Scott.
"Smart and Secure Cities and Communities."
Paper presented at Third International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership with Global City Teams Challenge (GCTC), Porto, Portugal. April 10, 2018 - April 13, 2018.

instead of establishing a new and separate "Security" Supercluster. This approach would have been easier, but much less likely to drive the greater security and data privacy awareness that is needed and must be sustained throughout GCTC communities and domain-specific projects.

To accomplish this goal, the 2018 GCTC-SC3 program will run a series of conferences during the course of 12 to 14 months, co-hosted by NIST and DHS S&T. The SC3 Kickoff conference, [8] the first in the series, will focus on (1) creating new Action Clusters that can demonstrate tangible benefits to the residents and citizens as well as considerations and readiness for cybersecurity and privacy and (2) seeding existing Action Clusters with the necessary expertise to enhance the security and privacy aspects of the project. It will be followed up by a mid-course "Tech Jam" event, which will feature presentations by the 2018 Action Clusters on their plans and progresses, and an Expo event in early 2019, which will demonstrate and share the accomplishments of all of the Action Clusters and SuperClusters. Throughout the 12-month process, the aim is to increase awareness of the importance of cybersecurity and privacy among smart city stakeholders, enabling cities and communities to identify, share and adopt replicable, secure, and privacy-enhancing best practices to address pressing issues they face every day.

Along with this approach, there are a number of particularly interesting topics and problems that can impact large parts of the entire GCTC effort, and will require NIST, DHS S&T and participants' attention. These include:

Mobile Systems Infrastructure: Most smart cities' efforts assume the presence of a reliable and increasingly capable mobility-based ecosystem, an ecosystem where the supporting network infrastructure includes greater vulnerability and risk than we should accept. As with electricity/power, how can we build the mobility ecosystem and support infrastructure towards the very high levels of reliability, assurance and resilience that are necessary in economies and societies built on these systems?

Test and Evaluation: Remembering that most complex systems development requires a near-perpetual cycle of assess, build, test/evaluate, and repeat -- how can we support the preparation, execution and learning from security and data privacy experiments, testing and evaluation? How can we do this across the wide variety of developmental efforts throughout the nation, and throughout the steady cycling of efforts year by year by year?

Capability Investment: As the nation's critical infrastructures continue their evolution and growth, how can we collaborate in connecting smart cities' security and data privacy efforts with smart national infrastructure investment throughout the nation? How do we invest in both the brawn and brains of infrastructure and critical infrastructure, throughout a diverse landscape?

Information Management: Enabled by comprehensive and sophisticated analytics, how do we effectively combine complex data analytics capabilities enabling successful smart cities operational and security outcomes – without overrunning necessary data privacy strategies, controls and checks that are absolutely necessary in our advanced democracy? How can we develop and apply a needed culture of trust as we experiment with these elements and tensions in emerging smart cities' systems?

Software Quality and Assurance: As all smart cities capabilities ultimately rely on software and software systems, how can we raise the quality of this key central element to every part of our smart cities and communities efforts? How can we connect open-source and proprietary, traditional IT software and mobile systems apps, legacy infrastructure and adaptive customer-facing software, all into a flexible, affordable and stronger quality overall outcome?

Business and Engagement Model: Most of the cities and communities are aware of the importance of cybersecurity and privacy to some extent. However, few of them are developing and executing a comprehensive and holistic plan to address the issues. It is partially due to the lack of resources and in-house expertise, but also the lack of mutually-beneficial engagement models between industry and municipal governments. What is the model that can incentivize industry and municipal governments to seriously engage on the issues of cybersecurity and privacy with high priority?

Education: How can we build technical and social learning elements spread throughout our smart cities efforts, to help users and citizens become active, stronger parts of the smart cities effort? How can we bring together operational learning systems coming out of particular smart cities efforts, with traditional/changing learning systems from educational sources increasingly focused on technical and social cyber-related learning?

V. Conclusion and Next Steps

The partnership between NIST and DHS S&T would bring the GCTC program to the next level. The partnership will not only introduce serious considerations of security and privacy to the mix, but also accelerate the adoption and deployment of replicable, secure, privacy-enhancing, and trustworthy technologies by cities and communities. The issues of security and privacy are critical cross-cutting elements for all relevant domains of smart cities – transportation, public safety, energy, water, broadband, and so on. In that sense, it is important to design in cybersecurity and privacy from the beginning of the development and deployment of solutions, and should not be added as an afterthought. The GCTC-SC3 is founded on the notion that the best approach to accelerate the adoption of security and privacy in smart city deployment is for the cities and communities to collaborate to identify, demonstrate, and replicate what works for them, in close partnership with technology providers. In the end, the success of GCTC-SC3 will be measured by the level of adoption of security and privacy considerations by Action Clusters and SuperClusters.

Decades ago, large parts of the military and industrial communities both needed performing at a higher level of

quality. History shows that since that time, the secret to industrial improvement has similarly been continuous innovation, learning and adaptation. "The Machine that Changed the World" shows the power of lean/adaptive/quality rather than mass production, "two very different ways of thinking about how humans work together to create value," and its eventual transition from manufacturing to other value-creating activities, from health care to retail to distribution. [9]

In parallel, the secret to substantial military improvement has been training, adaptation and continuous lessons-learned and self-examination. Senior military leaders acknowledge that the most significant military investment of the late 20th Century was not only equipment or personnel, but rigorous large-scale testing and evaluation of units and their leaders – in particular the National Training Center at Ft. Irwin and the Air Force "Red Flag" activities in Nevada.

Leaders in both these domains know that the secret to outstanding performance is continuously driving up quality through relentless organizational learning and improvement, and the nation needs this culture to emerge in smart city and community environments, for overall security, citizen data privacy and quality. We can encourage this through activities such as continuous experimentation; broad collaboration encouragement, support and leadership; and open collaboration and information sharing in a balance with separate competitive advantages.

Additionally, this must be both an Research and Development (R&D) and operational effort. Smart city and community efforts to build effective cybersecurity and data privacy capabilities cannot be accomplished when investigators, developers, engineers and operators are all working at arms-length from each other. Massachusetts Institute of Technology's Eric von Hippel describes a steady and inexorable shift away from classic central R&D organizations and sources, and towards more creative experimentation and innovation at the "edge" of modern, highly networked organizations. [10]

The NIST GCTC program has successfully embraced this approach, where both action clusters and area superclusters are working at the democratic edge of our complicated and complex nation of cities and communities. It is this distributed, innovative and independent collection of networked cities and communities that DHS S&T and NIST together are encouraging to go farther, and explore the best ways to develop cybersecurity and data privacy capabilities needed throughout our cities, communities, economy and nation.

DISCLAIMER

REFERENCES

[1] Rhee, S., "Catalyzing the Internet of Things and Smart Cities: Global City Teams Challenge," 2016 1st International Workshop on Science of Smart City Operations and Platforms Engineering in partnership with Global City Teams Challenge, Pages 1-4, Vienna, Austria, April 11, 2016, IEEE Explore Digital Library

[2] https://pages.nist.gov/GCTC/, last accessed on February 5, 2018

[3] Rhee, S., Burns, M., Nguyen, C., "Global City Teams Challenge 2016," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication. NIST-SP-1900-01, June, 2017. doi: 10.6028/NIST.SP.1900-01

[4] https://sloanreview.mit.edu/article/the-critical-difference-between-complex-and-complicated/, last accessed on February 5, 2018

[5] Griffor, E., Greer, C., Wollman, D., Burns, M., "Framework for cyber-physical systems: volume 1, overview," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication. NIST-SP-1500-201, June 26, 2017. doi: 10.6028/NIST.SP.1500-201

[6] https://www.nist.gov/cyberframework, last accessed on February 5, 2018

[7] https://pages.nist.gov/GCTC/super-clusters/, last accessed on February 5, 2018

[8] https://pages.nist.gov/GCTC/event/gctc-kickoff-2018/, last accessed on February 5, 2018

[9] Roos, D., Womack, J. P., Jones, D. T., "The Machine That Changed the World : The Story of Lean Production," Harper Perennial (November 1991), ISBN 0060974176, ISBN 978-0060974176

[10] Hippel, E., "Democratizing Innovation," The MIT Press (February 2005), ISBN 9780262002745, ISBN 9780262720472

Proceedings of the ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference IDETC/CIE 2018 August 26-29, 2018, Quebec City, Quebec, Canada

DETC2018-86055

A SUPER-METAMODELING FRAMEWORK TO OPTIMIZE SYSTEM PREDICTABILITY

Zhuo Yang, Douglas Eddy*, Sundar Krishnamurty, Ian Grosse
University of Massachusetts Amherst
Department of Mechanical and Industrial Engineering
Amherst, MA 01003
Email: [zhuoyang, dceddy]@engin.umass.edu
Email: [skrishna, grosse]@ecs.umass.edu

Yan Lu
National Institute of Standards and Technology
Engineering Laboratory
Gaithersburg, MD 20899
Email: yan.lu@nist.gov

ABSTRACT

Statistical metamodels can robustly predict manufacturing process and engineering systems design results. Various techniques, such as Kriging, polynomial regression, artificial neural network and others, are each best suited for different scenarios that can range across a design space. Thus, methods are needed to identify the most appropriate metamodel or model composite for a given problem. To account for pros and cons of different metamodeling techniques for a wide diversity of data sets, in this paper we introduce a super-metamodel optimization framework (SMOF) to improve overall prediction accuracy by integrating different metamodeling techniques without a need for additional data. The SMOF defines an iterative process first to construct multiple metamodels using different methods and then aggregate them into a weighted composite and finally optimize the super-metamodel through advanced sampling. The optimized super-metamodel can reduce an overall prediction error and sustains the performance regardless of dataset variation. To verify the method, we apply it to 24 test problems representing various scenarios. A case study conducted with additive manufacturing process data shows method effectiveness in practice.

KEYWORDS: predictive metamodeling, optimization

INTRODUCTION

Metamodeling techniques have been widely used to solve engineering design and optimization problems. A metamodel, also known as a surrogate model, uses statistics-based techniques to approximate the input-output relationship of a complex system, computer model or physical experiment without any knowledge of its internal structure [1]. It evolves from the classical Design of Experiments (DOE) theory [2]. Common metamodeling techniques include polynomial regression (PR) [3], Kriging [4], Radial Basis Function [5], support vector machine [6], and artificial neural network (ANN) [7, 8]. One technique may be superior to others for solving a specific problem, depending on the dimensionality, nonlinearity, sample size and sampling method of the problem [1, 9, 10]. Identifying an appropriate metamodeling technique is traditionally the first step of metamodel construction. A recent approach addressed the metamodel technique selection problem from the sampling perspective [11]. However, generic methods are still needed to address other factors that affect the selection of the best metamodeling technique.

Brute force search, also known as "generate-and-test", is the most general problem solving method that consists

*Corresponding author

Lu, Yan; Eddy, Douglas; Krishnamurty, Sundar; Grosse, Ian.
"A SUPER-METAMODELLING FRAMEWORK TO OPTIMIZE SYSTEM PREDICTABILITY."
Paper presented at ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2018) & Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. August 26, 2018 - August 29, 2018.

of systematically enumerating all possible candidate metamodeling techniques and selecting the most appropriate one based on a set of criteria [12]. This "generate-and-test" method is robust but lacks efficiency when the size of the candidate space is big. Modified methods were developed to improve the efficiency of exhaustive search using data-driven approaches [13–15]. Some research focuses on characterizing different metamodeling techniques to provide a metamodeling selection strategy. For example, Jin and coauthors suggest that PR should be implemented first to see if a reasonable fit can be obtained when constructing metamodels [9].

A selected technique is typically used for subsequent updates and predictions. However, data sampling can have a significant impact on the performance of a metamodeling technique [9]. Changing or adding data points could also affect the accuracy of a metamodel [1,16]. There is no guarantee that the technique with the lowest prediction error for one region in a design space will also have the lowest error for another region in that same space [17].

Integrated metamodeling approaches compose two or more techniques to address the issue of metamodel performance variation. For example, Turner (2005) introduced the Non-Uniform Rational B-splines method to model the hyperdimensional design spaces of products [18]. More conventionally, a method such as Universal Kriging combines PR and the ordinary Kriging method to improve modeling accuracy [4]. Grey-box metamodeling offers another approach to combine models of different fidelity levels [14,19]. Both universal Kriging and grey-box modeling techniques require two datasets to construct a metamodel. One is a high fidelity small dataset and the other a lower fidelity large dataset. However, it is not always possible in practice to generate additional customized data. Furthermore, only certain types of techniques can be integrated using the existing approaches. A more general method is thus needed that can combine any assortment of candidate metamodeling techniques to generate a composite that is best suited for a given problem globally. Such an approach should be broadly applicable and eliminate the need for case-by-case exploration for the best technique for every new data set.

This paper introduces a metamodel integration approach called the super-metamodel optimization framework (SMOF). In this introduction, SMOF integrates PR, Kriging, and ANN metamodels to improve accuracy over the individual metamodels. SMOF iteratively approaches the optimal combination of the techniques for a given dataset. The "Metamodeling Techniques" section summarizes the differences between the first set of candidate metamodeling techniques, including PR, Kriging and ANN. The "Super-metamodel Optimization Framework (SMOF)" section introduces the concept of SMOF. The "Test of SMOF Effectiveness" section presents the results from a designed case study, which compares the observations of SMOF performance under various modeling conditions. A manufacturing case study is also presented at the end of this section to prove its effectiveness in solving real world problems. Results are summarized and discussed in the "Discussion and Summary" section.

METAMODELING TECHNIQUES

This section summarizes the metamodeling techniques exemplified in this study. PR, Kriging and ANN techniques were selected because they span a diversity of parametric and non-parametric approaches. However, SMOF need not be limited to these algorithms. In general, any metamodeling technique could be considered as a candidate.

A metamodel can be expressed generally as [3]:

$$y(x) = \tilde{f}(x) + \varepsilon \tag{1}$$

Where $y(x)$ represents the relationship of an unknown system. $\tilde{f}(x)$ is an approximation of $y(x)$ derived statistically, and ε is the error term between the real and the approximated values, named residual. Here, x represents the set of the system's independent input variables. Different metamodeling techniques usually have different expressions of their approximation functions.

PR, Kriging and ANN methods were originally designed to solve different types of problems. PR metamodels can provide quick and accurate predictions for linear systems. Many researchers in engineering systems design have applied PR [20,21]. Kriging was initially developed for building highly nonlinear geological models [4]. The original goal of the ANN approach was to solve various problems in the same way that a human brain would [8]. However, none of these techniques have definite superiority over others in modeling different complex unknown systems.

To illustrate by an example, PR, Kriging and ANN were used to build three individual metamodels for a representative nonlinear function based on 200 Latin Hypercube Sampling (LHS) points [1,22]. The nonlinear function is given by Equation 2.

$$\begin{aligned} f(x_1,x_2) = 2 + 0.01(x_2 - x_1^2)^2 + (1-x_1) \\ +2(2-x_2)^2 + 7sin\left(\frac{x_1}{2}\right) \end{aligned} \tag{2}$$

The range of x_1 and x_2 is from -5 to 5. Forty additional randomly generated data points were used to validate the individual metamodels. Table 1 lists the average relative error magnitude (AREM) [16,23] of the three metamodels

Lu, Yan; Eddy, Douglas; Krishnamurty, Sundar; Grosse, Ian.
"A SUPER-METAMODELLING FRAMEWORK TO OPTIMIZE SYSTEM PREDICTABILITY."
Paper presented at ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2018) & Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. August 26, 2018 - August 29, 2018.

TABLE 1. Performance of individual metamodeling techniques

	PR	Kriging	ANN
AREM	22.5%	19.7%	12.0%
MAP (40 total points)	9	12	19

built with the same dataset. AREM is given by:

$$AREM = \frac{1}{m}\left(\frac{\sum_{i=1}^{m}|y_1 - \tilde{y}_i|}{y_i}\right) \quad (y_i \neq 0) \tag{3}$$

From Table 1, the ANN technique generates the lowest AREM, nearly half that of PR and Kriging. The ANN metamodel, however, does not always dominate the entire design space. The most accurate prediction (MAP) measures the number of validation points with lowest predictive error for each individual technique. Although the ANN metamodel has the lowest AREM, it has only 19 MAP out of the 40 total validation points. On the other hand, PR, the worst technique based on AREM, is superior at 9 validation points. Thus, combining three techniques across a design space could optimize prediction error.

SUPER-METAMODEL OPTIMIZATION FRAMEWORK (SMOF)

To account for pros and cons of different metamodeling techniques for a wide diversity of data sets, this section introduces the super-metamodel optimization framework (SMOF). Figure 1 shows the schematics of a SMOF model that is a composition of weighted individual models.

The following equations provide a general formulation of the SMOF model:

$$\tilde{f}(x) = \sum_{i=1}^{3} w_i \tilde{f}_i(x) \quad \left(\sum_{i=1}^{3} w_i = 1\right) \tag{4}$$

Where w_i is the weight factor of the i_{th} individual metamodel. Once the individual models are constructed, additional data can be used to find the optimal weight factors in the following procedure. The integrated SMOF model aims to minimize the sum of the prediction errors at these additional points, which is a function of the weight factors as expressed in the following equation.

$$Minimize: \sum_{j=1}^{n} Error^j(w_1, w_2, w_3) \tag{5}$$

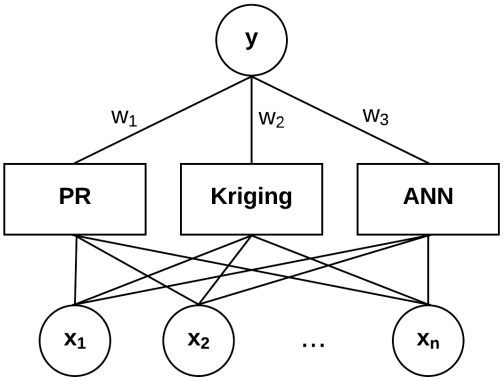


FIGURE 1. SMOF general model

where, $Error^j$ represents the relative prediction error of the weighted composed metamodel at $Point^j$. The absolute error of the composite model defined in Equation 4 can be calculated using Equation 3 for each data point. Equation 5 uniquely defines an optimization problem considering the performance of each individual metamodel at every data point. The resulting formulation of a relative $Error^j$ thereby follows in Equation 6.

$$\begin{aligned}
\textbf{Point}^{\mathbf{1}}: \quad \textbf{Error}^{\mathbf{1}} &= \left|\frac{w_1\tilde{y}_1^1 + w_2\tilde{y}_2^1 + w_3\tilde{y}_3^1 - y^1}{y^1}\right| \\
\textbf{Point}^{\mathbf{2}}: \quad \textbf{Error}^{\mathbf{2}} &= \left|\frac{w_1\tilde{y}_1^2 + w_2\tilde{y}_2^2 + w_3\tilde{y}_3^2 - y^2}{y^2}\right| \\
\textbf{Point}^{\mathbf{3}}: \quad \textbf{Error}^{\mathbf{3}} &= \left|\frac{w_1\tilde{y}_1^3 + w_2\tilde{y}_2^3 + w_3\tilde{y}_3^3 - y^3}{y^3}\right| \\
\vdots \qquad\quad \vdots \qquad\quad & \qquad\quad \vdots \\
\textbf{Point}^{\mathbf{j}}: \quad \textbf{Error}^{\mathbf{j}} &= \left|\frac{w_1\tilde{y}_1^j + w_2\tilde{y}_2^j + w_3\tilde{y}_3^j - y^j}{y^j}\right| \\
\vdots \qquad\quad \vdots \qquad\quad & \qquad\quad \vdots \\
\textbf{Point}^{\mathbf{n}}: \quad \textbf{Error}^{\mathbf{n}} &= \left|\frac{w_1\tilde{y}_1^n + w_2\tilde{y}_2^n + w_3\tilde{y}_3^n - y^n}{y^n}\right|
\end{aligned} \tag{6}$$

Where, y^j represents the observation value at the j_{th} point. $Error^j$ is calculated based on the weighted average of the predicted values from the individual models in Equation 4 and the observation value. After a transformation, the j_{th} point error becomes:

Lu, Yan; Eddy, Douglas; Krishnamurty, Sundar; Grosse, Ian.
"A SUPER-METAMODELLING FRAMEWORK TO OPTIMIZE SYSTEM PREDICTABILITY."
Paper presented at ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2018) & Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. August 26, 2018 - August 29, 2018.

$$\left|\frac{w_1\tilde{y}_1^j + w_2\tilde{y}_2^j + w_3\tilde{y}_3^j - y^j}{y^j}\right| = \left|\frac{\tilde{y}_1^j}{y^j}w_1 + \frac{\tilde{y}_2^j}{y^j}w_2 + \frac{\tilde{y}_3^j}{y^j}w_3 - 1\right| \quad (7)$$

Thus, the matrix formation of Equation 6 is:

$$\begin{bmatrix} \frac{\tilde{y}_1^1}{y^1} & \frac{\tilde{y}_2^1}{y^1} & \frac{\tilde{y}_3^1}{y^1} \\ \frac{\tilde{y}_1^2}{y^2} & \frac{\tilde{y}_2^2}{y^2} & \frac{\tilde{y}_3^2}{y^2} \\ \frac{\tilde{y}_1^3}{y^3} & \frac{\tilde{y}_2^3}{y^3} & \frac{\tilde{y}_3^3}{y^3} \\ \vdots & \vdots & \vdots \\ \frac{\tilde{y}_1^j}{y^j} & \frac{\tilde{y}_2^j}{y^j} & \frac{\tilde{y}_3^j}{y^j} \\ \vdots & \vdots & \vdots \\ \frac{\tilde{y}_1^n}{y^n} & \frac{\tilde{y}_2^n}{y^n} & \frac{\tilde{y}_3^n}{y^n} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad (8)$$

By applying linear algebra operations, a general optimization problem is formulated to minimize the sum of all prediction errors at all data points by choosing a set of weight factors. The weight factors are the independent variables to solve for in this optimization problem.

$$\begin{aligned} \textbf{Minimize}: \quad & \sum_{j=1}^{n}\sum_{i=1}^{3}\left|w_i\frac{\tilde{y}_i^j}{y^j} - 1\right| \\ \textbf{Subject to}: \quad & \sum_{i=1}^{3} w_i = 1 \\ & w_i \geq 0 \end{aligned} \quad (9)$$

Figure 2 introduces a SMOF-based work flow that targets to compose multiple metamodels into a super-metamodel to minimize the total prediction error as formulated in Equation 9. The preceding step in Figure 2 of "build individual models" is based on established principles to generate and test individual metamodels using a given dataset [12]. If additional data is available, a further step follows to derive a global optimal super-metamodel by composing the individual metamodels through a weight optimization procedure.

This SMOF-based approach also works when additional data is not available, by dividing the initial data set into three sub-sets: a set for metamodel construction, another set for super metamodel optimization, and the final set for metamodel validation. Since the optimal dataset for the original metamodel construction is unknown, it is necessary to iteratively segregate the data set until an acceptable AREM is achieved. As shown in Figure 2, the first step divides the given dataset into three sample sub-sets using the Minimum Euclidean Distance (MED) method [16]. Sub-set 1 is used to build individual metamodels. Sub-set 2 is used to find the optimal weight factors for the current super-metamodel. Sub-set 3 is used to verify whether the newly generated super-metamodel is better than the previous best super-metamodel and all the individual models.

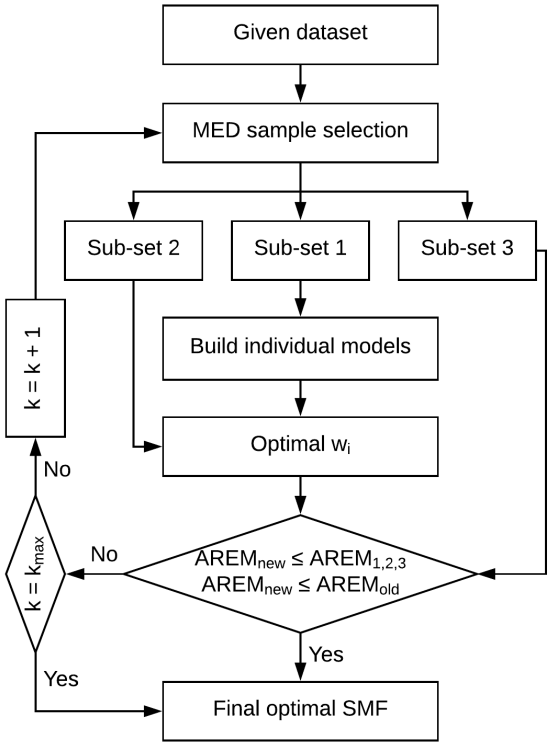


FIGURE 2. General procedure to build a SMOF metamodel

For our case study, the percentages of the 3 sample sub-sets are set to 60%, 20% and 20%, respectively. The final optimal super-metamodel will be obtained when the error criteria are met or the total number of iterations reaches its preset maximal, k_{max}. The final super-metamodel is composed as an optimally weighted sum of the last set of individual metamodels, as shown in Equation 4.

It is true that a cross-validation technique could be used to verify the metamodels. However, there is no guarantee that the metamodel with the lowest error from the training

Lu, Yan; Eddy, Douglas; Krishnamurty, Sundar; Grosse, Ian.
"A SUPER-METAMODELLING FRAMEWORK TO OPTIMIZE SYSTEM PREDICTABILITY."
Paper presented at ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2018) & Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. August 26, 2018 - August 29, 2018.

set will also have the lowest test error [17]. Thus, the SMOF procedure dedicates a separate data set of Sub-set 3 to verify whether or not the AREM criteria shown in Figure 2 are met.

TEST OF SMOF EFFECTIVENESS

This section presents two illustrative case studies to demonstrate the effectiveness of the SMOF method. The first case study applies SMOF to 3 different benchmark functions that have low, medium and high degrees of nonlinearity, respectively. Each function is further configured with various sample sizes and dimensionalities. The second case study tests the SMOF method using real manufacturing data. A Matlab build environment was deployed to execute the SMOF method and generate the results for these case studies. The ooDACE toolbox created the Kriging metamodels within Matlab [24].

Case Study 1: Benchmark Functions

To test the effectiveness of the SMOF method, 24 tests were designed to observe the hypothesized variations in recommended metamodeling techniques posed by differences in linearity, dimensionality, and sample size. Three benchmark functions with suggested different degrees of nonlinearity were deployed to generate the test datasets [25]. The Axis Parallel Hyper-Ellipsoid (APHE) function given below (continuous, convex and unimodal) generates data with a low-order of nonlinearity:

$$f(x) = \sum_{i=1}^{n} (i \cdot x_i^2) \qquad (-5.12 \le x_i \le 5.12) \tag{10}$$

The Rastrigin function (with frequent and regularly distributed local minima and multimodal) below generates data of medium-order nonlinearity:

$$f(x) = 10n + \sum_{i=1}^{n} [x_i^2 - 10\cos(2\pi x_i)] \quad (-5.12 \le x_i \le 5.12) \tag{11}$$

The Ackley function (with frequent local minima and highly multimodal) below generates high-order nonlinear data:

$$\begin{aligned} f(x) = \quad & -a \cdot \exp\left(-b \cdot \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i^2} \right) \\ & - \exp\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \cos(cx_i) \right) + a + \exp(1) \\ & -32.768 \le x_i \le 32.768 \\ & a = 20, \;\; b = 0.2, \;\; c = 2\pi \end{aligned} \tag{12}$$

TABLE 2. Design of experiment for case study 1

Problem No.	Nonlinearity	Sample Size	Dimensionality (# of variables)
1	Low	Small (20)	Small (n=2)
2	Low	Small (30)	Small (n=3)
3	Low	Small (50)	Medium (n=5)
4	Low	Small (80)	Large (n=8)
5	Low	Large (100)	Small (n=2)
6	Low	Large (300)	Small (n=3)
7	Low	Large (500)	Medium (n=5)
8	Low	Large (800)	Large (n=8)
9	Medium	Small (20)	Small (n=2)
10	Medium	Small (30)	Small (n=3)
11	Medium	Small (50)	Medium (n=5)
12	Medium	Small (80)	Large (n=8)
13	Medium	Large (100)	Small (n=2)
14	Medium	Large (300)	Small (n=3)
15	Medium	Large (500)	Medium (n=5)
16	Medium	Large (800)	Large (n=8)
17	High	Small (20)	Small (n=2)
18	High	Small (30)	Small (n=3)
19	High	Small (50)	Medium (n=5)
20	High	Small (80)	Large (n=8)
21	High	Large (100)	Small (n=2)
22	High	Large (300)	Small (n=3)
23	High	Large (500)	Medium (n=5)
24	High	Large (800)	Large (n=8)

The dimensionality variable was grouped into three categories: small scale (number of variables = 2 or 3), medium scale (number of variables = 5), and large scale (number of variables = 8). The LHS method [26] was used to generate the data for all of the problem configurations. To investigate the SMOF effectiveness for different sample sizes, two sampling scenarios were considered: small datasets with 10n samples (n represents the number of variables) and large datasets with 50n samples. Table 2 details the experimental design of the resulting 24 trials.

Lu, Yan; Eddy, Douglas; Krishnamurty, Sundar; Grosse, Ian.
"A SUPER-METAMODELLING FRAMEWORK TO OPTIMIZE SYSTEM PREDICTABILITY."
Paper presented at ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2018) & Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. August 26, 2018 - August 29, 2018.

TABLE 3. Test results of individual and SMOF metamodels

No.	PR	Kriging	ANN	SMOF	Weight
1	0.0000	0.0204	2.0429	0.0000	[1.0000, 0.0000, 0.0000]
2	0.0000	0.0600	0.3921	0.0000	[1.0000, 0.0000, 0.0000]
3	0.0000	0.2535	0.4661	0.0000	[1.0000, 0.0000, 0.0000]
4	0.0000	0.1914	0.3188	0.0000	[1.0000, 0.0000, 0.0000]
5	0.0000	0.0001	0.0698	0.0000	[1.0000, 0.0000, 0.0000]
6	0.0000	0.0001	0.0197	0.0000	[1.0000, 0.0000, 0.0000]
7	0.0000	0.0022	0.0629	0.0000	[1.0000, 0.0000, 0.0000]
8	0.0000	0.0181	0.0634	0.0000	[1.0000, 0.0000, 0.0000]
9	1.8630	1.1284	1.8787	1.1274	[0.0000, 0.9990, 0.0010]
10	0.5051	0.2699	0.5266	0.2298	[0.2143, 0.7796, 0.0061]
11	1.1268	0.1496	0.9306	0.1483	[0.1478, 0.8511, 0.0011]
12	1.3834	0.1727	0.2746	0.1712	[0.0001, 0.9771, 0.0228]
13	0.5209	0.0082	0.2540	0.0081	[0.0001, 0.9961, 0.0038]
14	0.5545	0.3425	0.4110	0.3218	[0.0001, 0.9163, 0.0836]
15	0.3489	0.2164	0.4519	0.2105	[0.0867, 0.9115, 0.0018]
16	0.2649	0.2363	0.2523	0.2348	[0.0001, 0.7392, 0.2607]
17	0.1047	0.0714	0.1727	0.0564	[0.3402, 0.6570, 0.0028]
18	0.1251	0.1485	0.2554	0.1160	[0.5732, 0.1450, 0.2818]
19	0.0794	0.0836	0.1233	0.0255	[0.5068, 0.3726, 0.1206]
20	0.0548	0.1025	0.1657	0.0357	[0.5904, 0.2778, 0.1318]
21	0.0715	0.0943	0.1255	0.0608	[0.4986, 0.1013, 0.4001]
22	0.0459	0.0857	0.0825	0.0380	[0.5948, 0.1549, 0.2503]
23	0.0377	0.0490	0.0468	0.0287	[0.6739, 0.2075, 0.1186]
24	0.0258	0.0364	0.0436	0.0238	[0.6287, 0.3384, 0.0329]

Lu, Yan; Eddy, Douglas; Krishnamurty, Sundar; Grosse, Ian.
"A SUPER-METAMODELLING FRAMEWORK TO OPTIMIZE SYSTEM PREDICTABILITY."
Paper presented at ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2018) & Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. August 26, 2018 - August 29, 2018.

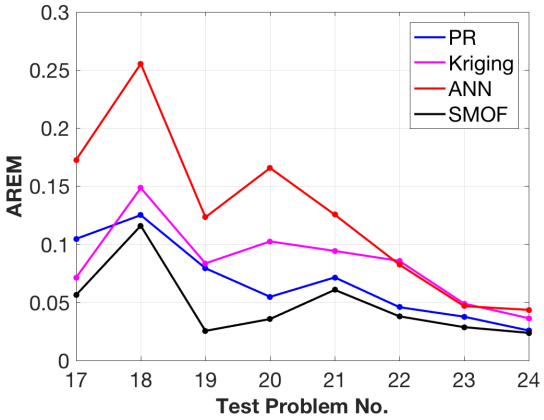


FIGURE 3. AREM of PR, Kriging, ANN, and SMOF metamodels for test problems 17 to 24 with high-order nonlinearity

Table 3 summarizes the results of running these trials. The initial AREM was set to a very large value to enable iterations. The maximum number of iterations was set to 100 for all trails. For the more linear data generated by trials 1 through 8, the PR technique perfectly dominates the entire design space with prediction errors close to 0. Accordingly, the super-metamodel generates exactly the same AREM as the PR metamodel does, consistent with the optimal weights of [1, 0, 0]. For trials 9 to 16, the Kriging metamodel performed better than the other two techniques. Therefore, it weighs heavily on the super-metamodel and has comparable AREMs.

The highly nonlinear trials of 17 through 24 reveal more about the effects of sample size and dimensionality on the SMOF metamodel. The AREMs of the individual metamodels and the SMOF super-metamodels are shown in Figure 3. For all of the 8 trials, none of the individual metamodels is always superior to the other two and the SMOF super-metamodel outperforms all of the three individual metamodels. Table 3 shows that even the least accurate individual metamodel has a nonzero weight factor in the super-metamodel composition. This indicates that every individual metamodel contributes to the super-metamodel accuracy for this type of function with higher-order nonlinearity. Thus, Table 3 suggests that the SMOF is most beneficial for highly nonlinear problems. These results in Table 3 also indicate that use of the SMOF may reveal insights about the order of linearity of a given data set.

The test result in Table 3 was calculated from Sub-set 3, which was segregated from the original given data. In this case, additional data points can be generated to further verify the effectiveness of the SMOF method. For the following study, additional 250n data points (n = number of independent variables) were generated using the Monte Carlo method.

TABLE 4. AREM results from additional data

No.	PR	Kriging	ANN	SMOF
1	0.0000	0.0115	13.4318	0.0000
2	0.0000	0.3016	0.9600	0.0000
3	0.0000	0.4044	0.6045	0.0000
4	0.0000	0.2553	0.3807	0.0000
5	0.0000	0.0000	0.0831	0.0000
6	0.0000	0.0000	0.0342	0.0000
7	0.0000	0.0011	0.1577	0.0000
8	0.0000	0.0147	0.2042	0.0000
9	1.8010	1.3374	1.5044	1.3359
10	0.6435	0.5986	0.6545	0.5866
11	0.4357	0.3453	0.4740	0.3439
12	2.0602	0.2498	0.3157	0.2420
13	1.4063	0.0147	0.1255	0.0146
14	0.5357	0.2516	0.3942	0.2453
15	0.3720	0.2973	0.3808	0.2859
16	0.2337	0.2197	0.2603	0.2087
17	0.1214	0.0989	0.2251	0.0896
18	0.1024	0.1396	0.2220	0.0959
19	0.0974	0.1130	0.1274	0.0709
20	0.1281	0.0904	0.1282	0.0894
21	0.1089	0.1056	0.1126	0.0950
22	0.0679	0.0643	0.0592	0.0551
23	0.0422	0.0478	0.0583	0.0367
24	0.0288	0.0392	0.0456	0.0275

The results for these trials are given in Table 4. For a consistent comparison, none of the existing metamodels were updated using any of this additional data. Therefore, the weight factors for all of the trials remain the same as shown in Table 3. The formerly developed SMOF super-metamodels still consistently outperform all the individual metamodels as shown in Figure 4, even though for some cases, for example in trial No. 20, the performance ranks of PR and Kriging have switched.

Lu, Yan; Eddy, Douglas; Krishnamurty, Sundar; Grosse, Ian.
"A SUPER-METAMODELLING FRAMEWORK TO OPTIMIZE SYSTEM PREDICTABILITY."
Paper presented at ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2018) & Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. August 26, 2018 - August 29, 2018.

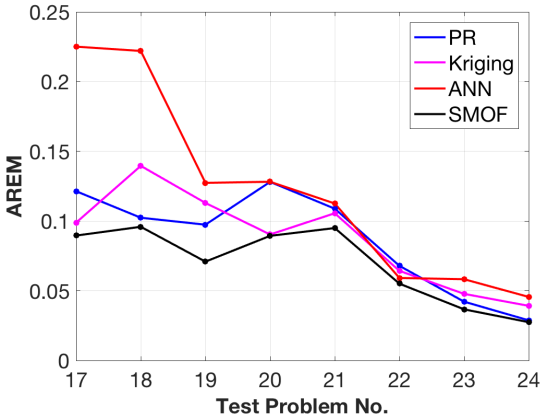


FIGURE 4. AREM of PR, Kriging, ANN, and SMOF models for test problems 17 to 24 using additional data

TABLE 5. Variable values in experiment

Variable	Values
Scan speed (mms^{-1})	50, 100, 200, 300, 400, 500
Scan spacing (μm)	25, 50, 75
Pulse frequency (kHz)	0, 10, 20, 30, 40, 50, 60

Case Study 2: A Manufacturing Application

This section illustrates a test of the SMOF effectiveness with real data from an additive manufacturing process. This example utilizes an experimental dataset of a Direct Metal Laser Re-Melting (DMLRM) additive manufacturing process [27]. DMLRM is a process variant of selective laser sintering (SLS). The experiment studied the effects of scanning speed, scanning spacing and laser pulse frequency on the relative density of the parts produced by this SLS process. The experimental design was based on a fractional factorial DOE with 105 total trials for the three input variables with values sets as shown in Table 5. The laser power remained fixed at 80 W for all the experiments.

To implement SMOF, the data was first divided into 3 sub-sets using MED sampling method at every iteration, as shown in Figure 2. 63 data points (60%) were included in Sub-set 1 to build individual metamodels. Sub-set 2 and Sub-set 3 each include 21 different data points (20%). At the end of a total of 100 iterations, the weight factor vector is optimized to $w = [0.151, 0.571, 0.277]$. Table 6 shows the superiority of the SMOF super-metamodel compared to the three individual metamodels in this case. These results also verify that the individual metamodel with the smallest error always has the largest weight value. Based on the discussions of the prior subsection, the results in Table 6 also indicate that this problem likely has a high level of nonlinearity.

TABLE 6. AREM for individual and SMOF metamodels, and corresponding weight factors

	PR	Kriging	ANN	SMOF
Optimal Weight	0.151	0.571	0.277	N/A
Final AREM	9.50%	6.72%	9.42%	5.47%

DISCUSSION AND SUMMARY

The main objective of this work is to compose super-metamodels from multiple metamodeling techniques for a better global accuracy without the need to generate additional data, which can be expensive. The idea of the SMOF introduces several salient features. First, a matrix of all data points and metamodeling techniques accounts for each corresponding error value. Second, a composite weighted formulation aggregates predicted values from various metamodeling techniques. Third, iterative sampling of the data optimizes error as a function of the weights' vector to find the optimal weighted composite for a given problem.

These innovations lead to several main benefits. First, the advantages and disadvantages of different techniques for various conditions of different types of problems become irrelevant. Thus, superior predictive accuracy can be assured regardless of which technique is best for a given problem. Second, use of the SMOF provides some indication about the degree of linearity of the problem. Third, it becomes unnecessary to generate expensive additional data to overcome any inaccuracies related to technique selection. These benefits should help to address uncertainty about which metamodeling technique to use for a given problem, which should result in more consistent predictive accuracy.

All the case studies in the previous section corroborate the hypothesis that the proposed SMOF approach of an iteratively optimized and weighted composite of individual techniques can significantly and consistently improve prediction accuracy over individual techniques regardless of the sample size and the dimensionality of the data. The first case study also strongly suggests that significance of these advantages can increase with the degree of nonlinearity of a dataset. However, even for relatively linear conditions, SMOF can simultaneously verify the degree of linearity and reveal which metamodeling technique is best to use for that given dataset. Although only the PR, Kriging and ANN techniques were used in this study, the overall approach should be extendable to a mixture of other metamodeling techniques.

Lu, Yan; Eddy, Douglas; Krishnamurty, Sundar; Grosse, Ian.
"A SUPER-METAMODELLING FRAMEWORK TO OPTIMIZE SYSTEM PREDICTABILITY."
Paper presented at ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2018) & Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. August 26, 2018 - August 29, 2018.

There are several notable limitations of the SMOF method to address. No more than 800 data points could verify the method for the data sets used in this work. Future work could examine scenarios of more complicated conditions such as large data sets of more than 10,000 data points. These larger datasets could run with various resolutions, and with various weighting schemes of the points in the design space to test more comprehensively.

The iterative process does require more computation than the established exhaustive "generate-and-test" approach [12], which is identical to the first iteration of SMOF execution without any optimization. Thus, users should be aware of this inherent tradeoff between predictive accuracy and computational cost, especially for large datasets on a case-by-case basis. Computations of the SMOF approach could be reduced by partially updating the sample sets for each iteration. The second limitation concerns the opportunity to improve the predictive accuracy of the individual metamodels. The current data sub-set segregation by the MED method runs the same sampling process at each iteration. However, it does not necessarily optimize all the metamodels. A second sampling stage of sequential infilling techniques could be introduced to make improvements. Thus, a research opportunity exists for novel sampling techniques that could further improve both predictive accuracy and computational efficiency.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1439683, the National Institute of Standards and Technology (NIST) under Cooperative Agreement number NIST 70NANB15H320, and industry members of the NSF Center for e-Design.

REFERENCES

[1] Shao, T., and Krishnamurty, S., 2008. "A clustering-based surrogate model updating approach to simulation-based engineering design". Journal of Mechanical Design, 130(4), p. 041101.

[2] Wang, G. G., and Shan, S., 2007. "Review of metamodeling techniques in support of engineering design optimization". Journal of Mechanical design, 129(4), pp. 370–380.

[3] Box, G. E., and Draper, N. R., 1987. Empirical model-building and response surfaces. John Wiley & Sons.

[4] Cressie, N., 2015. Statistics for spatial data. John Wiley & Sons.

[5] Dyn, N., Levin, D., and Rippa, S., 1986. "Numerical procedures for surface fitting of scattered data by radial functions". SIAM Journal on Scientific and Statistical Computing, 7(2), pp. 639–659.

[6] Vapnik, V., 2013. The nature of statistical learning theory. Springer science & business media.

[7] Rosenblatt, F., 1958. "The perceptron: a probabilistic model for information storage and organization in the brain.". Psychological review, 65(6), p. 386.

[8] Yegnanarayana, B., 2009. Artificial neural networks. PHI Learning Pvt. Ltd.

[9] Jin, R., Chen, W., and Simpson, T. W., 2001. "Comparative studies of metamodelling techniques under multiple modelling criteria". Structural and multidisciplinary optimization, 23(1), pp. 1–13.

[10] Simpson, T., Mistree, F., Korte, J., and Mauery, T., 1998. "Comparison of response surface and kriging models for multidisciplinary design optimization". In 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, p. 4755.

[11] Garbo, A., and German, B., 2017. "Adaptive sampling with adaptive surrogate model selection for computer experiment applications". In 18th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, p. 4430.

[12] Narendra, P. M., and Fukunaga, K., 1977. "A branch and bound algorithm for feature subset selection". IEEE Transactions on computers, 9(C-26), pp. 917–922.

[13] Rice, J. R., 1976. "The algorithm selection problem". In Advances in computers, Vol. 15. Elsevier, pp. 65–118.

[14] Cui, C., 2016. Building Energy Modeling: A Data-Driven Approach. Arizona State University.

[15] Yang, Z., Hagedorn, T., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., Lu, Y., and Witherell, P., 2017. "A domain-driven approach to metamodeling in additive manufacturing". In ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V001T02A028–V001T02A028.

[16] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., and Lopez, F., 2016. "Investigating predictive metamodeling for additive manufacturing". In ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V01AT02A020–V01AT02A020.

[17] James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013. An introduction to statistical learning, Vol. 112. Springer.

[18] Turner, C. J., 2005. "Hypermodels: hyperdimensional performance models for engineering design". PhD the-

Lu, Yan; Eddy, Douglas; Krishnamurty, Sundar; Grosse, Ian.
"A SUPER-METAMODELLING FRAMEWORK TO OPTIMIZE SYSTEM PREDICTABILITY."
Paper presented at ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2018) & Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. August 26, 2018 - August 29, 2018.

sis.
[19] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., Lu, Y., and Witherell, P., 2017. "Investigating grey-box modeling for predictive analytics in smart manufacturing". In ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V02BT03A024–V02BT03A024.
[20] Unal, R., Lepsch, R., Engelund, W., and Stanley, D., 1996. "Approximation model building and multidisciplinary design optimization using response surface methods". In 6th Symposium on Multidisciplinary Analysis and Optimization, p. 4044.
[21] Engelund, W. C., Stanley, D. O., Lepsch, R. A., McMillin, M. M., and Unal, R., 1993. "Aerodynamic configuration design using response surface methodology analysis". NASA STI/Recon Technical Report A, 94.
[22] Martin, J. D., 2009. "Computational improvements to estimating kriging metamodel parameters". Journal of Mechanical Design, 131(8), p. 084501.
[23] Shao, T., 2007. Toward a structured approach to simulation-based engineering design under uncertainty. University of Massachusetts Amherst.
[24] Couckuyt, I., Dhaene, T., and Demeester, P., 2014. "oodace toolbox: a flexible object-oriented kriging implementation". Journal of Machine Learning Research, 15, pp. 3183–3186.
[25] Tang, K., Yáo, X., Suganthan, P. N., MacNish, C., Chen, Y.-P., Chen, C.-M., and Yang, Z., 2007. "Benchmark functions for the cec'2008 special session and competition on large scale global optimization". Nature Inspired Computation and Applications Laboratory, USTC, China, 24.
[26] McKay, M. D., Beckman, R. J., and Conover, W. J., 2000. "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code". Technometrics, 42(1), pp. 55–61.
[27] Morgan, R., Sutcliffe, C., and O'neill, W., 2004. "Density analysis of direct metal laser re-melted 316l stainless steel cubic primitives". Journal of materials science, 39(4), pp. 1195–1205.

Lu, Yan; Eddy, Douglas; Krishnamurty, Sundar; Grosse, Ian.
"A SUPER-METAMODELLING FRAMEWORK TO OPTIMIZE SYSTEM PREDICTABILITY."
Paper presented at ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2018) & Computers and Information in Engineering Conference, Quebec City, Quebec, Canada. August 26, 2018 - August 29, 2018.

Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference
IDETC/CIE 2018
August 26-28, 2018, Quebec City, Canada

DETC2018-85996

SELF-IMPROVING ADDITIVE MANUFACTURING KNOWLEDGE MANAGEMENT

Yan Lu
Systems Integration Division
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
Email: yan.lu@nist.gov

Zhuo Yang, Douglas Eddy, Sundar Krishnamurty
University of Massachusetts Amherst
Department of Mechanical and Industrial Engineering
Amherst, MA 01003
Email: [zhuoyang, dceddy]@engin.umass.edu, skrishna@ecs.umass.edu

ABSTRACT

The current additive manufacturing (AM) product development environment is far from being mature. Both software applications and workflow management tools are very limited due to the lack of knowledge supporting engineering decision making. AM knowledge includes design rules, operation guidance, and predictive models, etc., which play a critical role in the development of AM products, from the selection of a process and material, lattice and support structure design, process parameter optimization to in-situ process control, part qualification and even material development. At the same time, massive AM simulation and experimental data sets are being accumulated, stored, and processed by the AM community. This paper proposes a four-tier framework for self-improving additive manufacturing knowledge management, which defines two processes: bottom-up data-driven knowledge engineering and top-down goal-oriented active data generation. The processes are running in parallel and connected by users, therefore forming a closed loop, through which AM knowledge can evolve continuously and in an automated way.

Keywords: additive manufacturing, knowledge management, manufacturing system integration

1. INTRODUCTION

Many hurdles continue to hinder the widespread adoption of additive manufacturing (AM) technologies for production, including low repeatability and quality inconsistency, high cost and time in qualification, and constrained material choices [1]. A key step in overcoming each of these hurdles is to obtain the knowledge needed to support engineering decision making through the AM product development lifecycle and across its value chain. In recent years, through the acquisition of "know how" startups, major CAD/CAM vendors have quickly expanded their offerings to include reverse engineering, geometry repairing, topology optimization, build preparation and process simulation in support of "Design for AM" engineering activities [2]. There are also services like 3D Hubs where engineers can get instant feedback on part manufacturability and the best processes for the design [3].

However, existing AM software know-how is still far from being mature enough to allow engineers to fully grasp the requirements and limitations to bring optimized AM parts to production [4]. Albright [5] lists some missing capabilities from the existing AM software, including design rules to validate issues of the minimum wall thickness, printability of the part overhang angle, shrinkage/warping prediction, support design, orientation selection, lattice structure analysis and post-process planning, etc. These desired but lacking software functions demand a new set of knowledge about AM capabilities, limitations and design rationales, which not only depends on the choice of material and technology but is also determined by the product definition and process parameters. Both today's design and manufacturing engineers and university graduates are encouraged to seek out the knowledge associated with the 'physics' of how AM processes work and how within each process category each type of material may respond differently

Lu, Yan; Yang, Zhuo; Eddy, Douglas; Krishnamurty, Sundar.
"Self-Improving Additive Manufacturing Knowledge Management."
Paper presented at Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, Quebec City, Canada. August 26, 2018 - August 28, 2018.

when trying to build a specific geometry in a specific orientation [4].

While some researchers are working on physics-based modeling and simulation techniques to understand AM processes, others are diligently conducting field studies and disseminating information to derive process-structure-property (PSP) relationships directly from data. Data-driven modeling and information fusion are foundational to AM knowledge development [6]. Such approaches take experimental or simulation data, using advanced data analytics techniques, such as metamodeling and machine learning, to identify AM material PSP relationships and derive AM design, process planning, and operation rules. Illustrated as a bottom-up process in Figure 1, new data sets emerge first, whether from research experiments in the lab or real production on the manufacturing shop floor, leading to new information fused and analyzed, so that new AM knowledge is acquired, and AM software functions are enhanced.

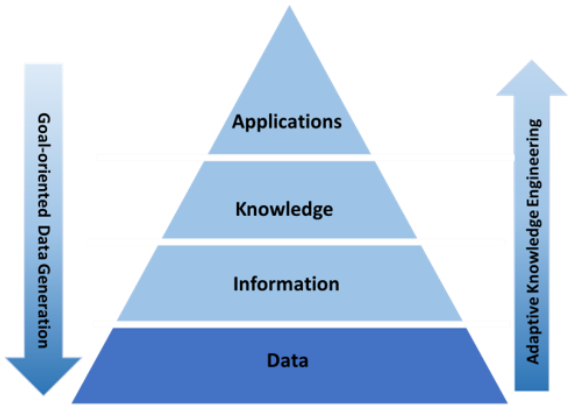


Figure 1: A modified Data-Information-Knowledge-Wisdom (DIKW) model

Many individual techniques have been studied and reported to support data-driven AM knowledge engineering. Kim et al. gave a comprehensive description of the types of heterogeneous data sets generated and consumed during an AM development lifecycle [7]. Lu et al. have reported a common information model and a collaborative database to structure and fuse the heterogeneous data sets contributed by different stakeholders in the AM community [8, 9]. Towards advanced data analytics, intensive research has been done by Yang et al in employing metamodeling techniques in design and optimization [10, 11]. AM builds are well suited for the reported collaborative data and information management in [9], especially for metal parts, where costs for conducting large scale sampling over many variables can be prohibitive. Generative learning and transfer learning are two methods reported in other domains, which might be used for AM metamodeling based on heterogeneous data sets.

At the same time, a few research activities were reported related to the techniques required for an application-driven data generation process, shown in Figure 1 as a top-down process. An Information Fusion Enterprise Model proposed by Kessler and White provides an approach about how to derive information needs from user's queries [12]. Adaptive sampling or sequential sampling has been used widely for metamodel improvement through new data acquisition.

The top-down and bottom-up processes are running in parallel and isolated without the AM community being aware of the opportunity and benefit to integrate and streamline them. The existing research works surveyed above only address the individual functions and links of the processes without a vision of creating an integrated workflow. Current disconnected AM knowledge management makes it harder for AM engineers to fully grasp the benefits and limitations of AM technology and bring optimized AM designs to production. In this paper, we proposed a self-improving additive manufacturing knowledge management approach which consists of a bottom-up data-driven knowledge engineering process and a top-down goal-oriented active data generation process and forms a closed-loop for continuous knowledge improvement. The proposed approach is based on a four-tier data-information-knowledge-wisdom (DIKW) model variant as shown Figure 1. The original DIKW model [13] is modified to have the top layer renamed as "Applications" to make it more understandable for AM engineers. Besides, a bottom-up and a top-down process are added to the pyramid to capture the need of workflow integration and automation for the proposed self-improving knowledge management. The bottom-up process is named as "Adaptive Knowledge Engineering" while the top-down one is called "Goal-oriented Data Generation".

The paper is organized as follows. Section 2 introduces the layer model for the four-tiers knowledge engineering framework. Section 3 describes the top-down and bottom-up processes and how they are connected into a closed loop. Section 4 provides an example to show how a self-improving AM knowledge management system works. Section 5 summarizes the paper and discusses our future work.

2. A FRAMEWORK FOR SELF-IMPROVING ADDITIVE MANUFACTURING KNOWLEDGE MANAGEMENT

An elaborated four-tier knowledge management framework for AM is shown in Figure 2. The Data layer sits at the bottom, captures diverse data sets generated and used in AM lifecycle and value chains. The Information tier fuses the heterogeneous data sets and manages them in a collaborative way. The Knowledge layer sits on top of the information layer, capturing process, machine and material capabilities, design rules, operating guidance, process models and asset health models

Lu, Yan; Yang, Zhuo; Eddy, Douglas; Krishnamurty, Sundar.
"Self-Improving Additive Manufacturing Knowledge Management."
Paper presented at Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, Quebec City, Canada. August 26, 2018 - August 28, 2018.

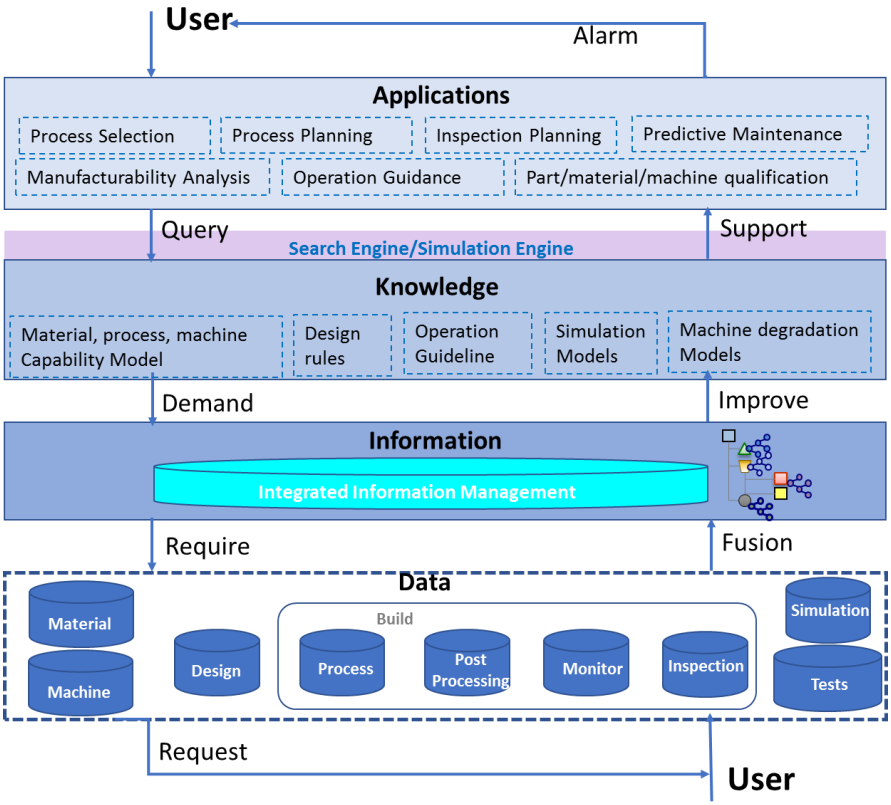


Figure 2: A 4-tier analytical framework for self-improving AM knowledge management

which can be queried and simulated using various search engines and simulation engines. The top tier is the Applications layer, which consists of software applications in support of AM lifecycle and value chain activities. Knowledge is the key for AM engineers to conduct their activities and make correct design or operation decisions. A summary of the components at each layer is provided below.

Data Layer:

The data components are heterogeneous, covering vendor provided machine and material data, asset and feedstock data from their owners, design data from designers, process data from manufacturers, test data from inspectors and all kinds of experimental data from AM researchers. They are summarized in Table 1.

Information Layer:

A common conceptual information model captures the data sets generated from AM lifecycle and value chain, as shown in Figure 3 [7]. Based on the common data model, heterogeneous data sets are fused into a collaborative information system based on NoSQL technology for both easy query and efficient storage [8].

Table 1: AM data types summary

Data Category	Data Description
Material	Material type and grades; Vendor provided material properties (feedstock and as-built); Material stock information and actual material properties
Machine	Process type; Vendor provided machine specifications; Machine information as an asset; asset maintenance information
Design	CAD models; Design meta data; Design intents and PMIs; design features
Build	Build meta information; Feedstock material information; Equipment information; Structure and support as designed; Process parameters; Preprocess pedigree; In-situ monitoring data; Post processing information; Inspection data
Tests	Test meta information; Sample information; Test type/standards; Operator information; Test results.
Simulation	Simulation models; simulation configurations; simulation results

Lu, Yan; Yang, Zhuo; Eddy, Douglas; Krishnamurty, Sundar.
"Self-Improving Additive Manufacturing Knowledge Management."
Paper presented at Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, Quebec City, Canada. August 26, 2018 - August 28, 2018.

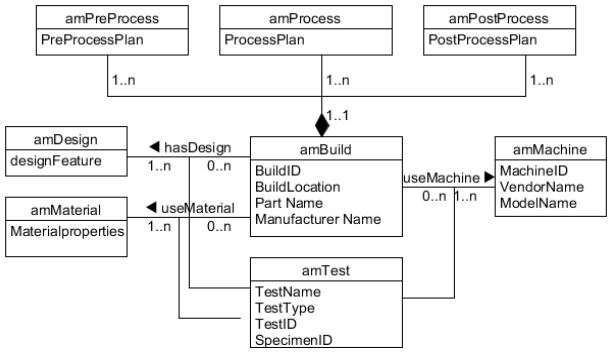


Figure 3: An AM common information Model [8]

Knowledge Layer:

AM knowledge can be classified into two categories: descriptive and prescriptive. The descriptive knowledge could be either physics-based models or metamodels, both of which can be used to simulate AM processes and allow engineers to perform *'what if'* scenarios for potential parametric optimization. MathML [14] and PMML [15] are two markup languages frequently used to represent and communicate the models.

The prescriptive knowledge includes design rules, operation guidance and diagnosis rules, which can be applied directly to AM applications. Dedicated AM design rules that relate to process capabilities are necessary for both CAD tools and AM process planning tools. Design rules can be represented and executed using an ontology. Figure 4 shows an ontology that can be used to select the best manufacturing process, considering AM as an option to compare, for a given part [16],17].

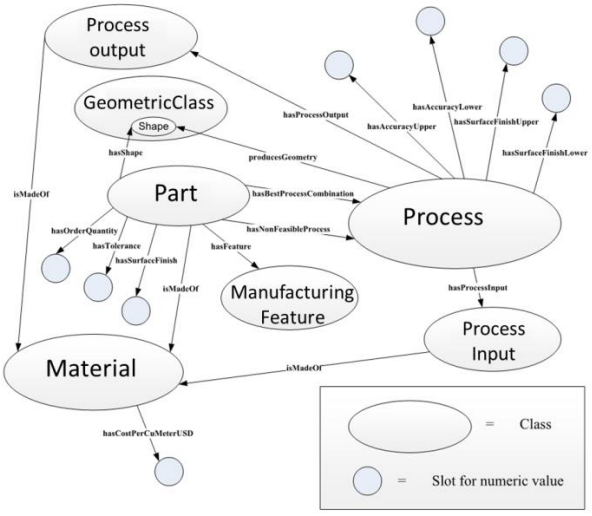


Figure 4: An Ontology for manufacturing process selection [16],17]

Applications Layer:

This layer captures software applications that support end-to-end digital processing during the AM product lifecycle and across its value chain. The software can be hosted on clouds and provided as services to AM stakeholders. Aided by effective workflow management, compositions of software functions can greatly streamline how AM products are designed, manufactured and tested. Figure 5 shows a list of AM applications hosted on a collaborative development platform supported by a shared knowledge base.

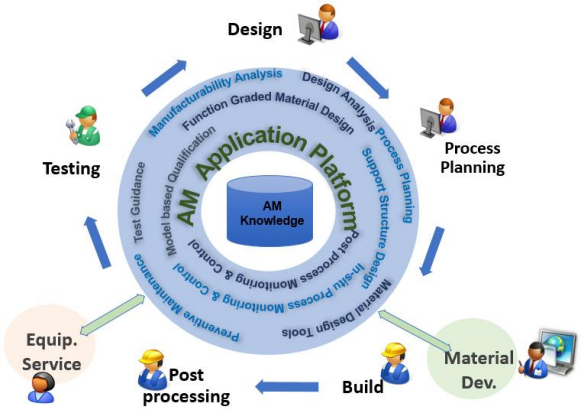


Figure 5: A collaborative development platform for AM

3. PARALLEL PROCESSES FOR SELF-EVOLVING AM KNOWLEDGE MANAGEMENT

Our previous work has been focused on how to identify, model and represent the data, information, and knowledge for each layer of the framework [6]. This section introduces our latest work on streamlining the process of engineering AM knowledge from data. In addition, a complementary top-down process is introduced to prescribe data sets to accelerate AM knowledge accumulation and enable the self-evolving AM knowledge management. Requirements on individual technologies are identified corresponding to those links between the layers of the tiered framework as well as the connections between the two-way processes.

3.1 Bottom-up Process

The bottom-up process, Adaptive Knowledge Engineering, consists of multiple sub-processes including heterogonous data generation and curation, data integration and information fusion, knowledge extraction and fusion, and knowledge query and access. Substantial research has been conducted on the sub-processes individually. The focus of our framework is to provide a method to automate the data-driven knowledge engineering process and allow for a continuously improved knowledge management system.

Figure 6 illustrates a typical adaptive knowledge engineering workflow. It starts with a new data set being available for

Lu, Yan; Yang, Zhuo; Eddy, Douglas; Krishnamurty, Sundar.
"Self-Improving Additive Manufacturing Knowledge Management."
Paper presented at Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, Quebec City, Canada. August 26, 2018 - August 28, 2018.

knowledge engineering, which kicks off a sequential workflow consisting of sub-processes to extract knowledge out of the data sets and fuse it with the existing knowledge.

Step 1: Check the data trustworthiness and quality; after the data source is verified and the data quality is validated, curate the data set.

Step 2: Check if the new data set brings in any new information, for example, a new type of material, a new model of AM machine or a new build, etc. If the answer is affirmative, fuse the new information into the existing information system. Information fusion could involve feature recognition to characterize a part design when a new set of build data is ingested. In addition, the links between the build and the material/machine products used for such a build, if already captured, should be established.

Step 3: Check if the new information leads to any knowledge update. If yes, extract knowledge from the new information and update the knowledge base. For example, minimum thin wall thickness is a measure characterizing the capability of an AM machine. This capability can be assessed based on a statistical analysis of all the builds made on the machine. Therefore, a new build will likely update that knowledge of the machine capability. Similarly, existing predictive models or design rules for a material-process combination can be updated with new information.

Step 4: Check if any ongoing design decisions are still valid after the knowledge base update. If not, re-initiate the affected design processes. If yes, actions will be taken into the corresponding AM activities, and new data sets are expected.

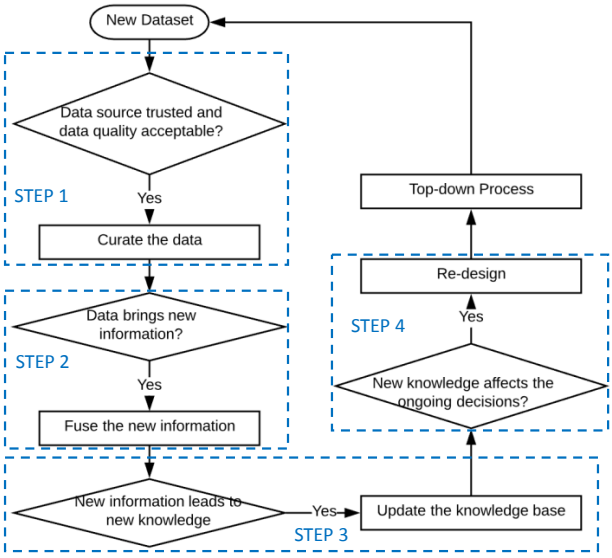


Figure 6: A typical bottom-up process

Various data integration approaches have been proposed to enable the automation of Step 1 and Step 2 of the bottom-up process. At NIST, we proposed to use a common information model to ingest diverse data sets, which is more suitable for greenfield data integration. If legacy databases exist, an ontology based data integration will be more appropriate [18].

Automation of Step 3 is still a very open problem. The challenges lie in today's immature model/knowledge characterization and representations methods [19,20]. Both rules and models in the knowledge base should be characterized for their validity and applicability in order to evaluate if a certain piece of information will enrich the knowledge or not. However, strong dependencies of AM process outputs on part geometry make the standardization and representation of AM design rules very challenging to automate. The next section takes a case study to illustrate how metamodels can be improved with new data sets. The final step can be easily realized through a publish/subscribe software implementation, where AM applications are programmed to receive knowledge updates and check the consistency between the ongoing decisions and the newly updated knowledge.

3.2 Top-down Process

Individual experimental studies often contain only a few measurements and focus on specific sub-processes. Costs for conducting large scale sampling over many process variables can be prohibitive for AM data generation. However, the small sets of data captured do not adequately represent the inherently large sets of process variables and material microstructure variances that must be analyzed for establishing AM PSP relationships. Therefore, it is critical for the AM community to work jointly in a coordinated and systematic fashion to generate data sets which can maximize AM knowledge discovery. The top-down process is designed to solve an optimal data generation problem based on a goal-oriented method.

As shown in Figure 7, the top-down process starts when a query is made in an AM application for knowledge to support engineering decision making. Five steps are involved in completing an optimal data set generation process.

Step 1: Query for design rules for design decisions or a request for simulations to conduct "what-if" analysis.

Step 2: Check if the requested design rules or simulation models exist in the knowledge. If not, define the information necessary to derive such rules or models and issue a query for the information.

Step 3: Check if the information already exists in the information system. If not, identify the various data sets needed for the information formation.

Step 4: Check if all the data sets already exist in distributed data sources. If not, call for the design of experiments (DOE) and data contribution.

Step 5: DOE is conducted, and the list of experiments is distributed to the AM community. Individual data contributors conduct the experiments and make new data set submissions, which kicks off a new round of the bottom-up process.

Lu, Yan; Yang, Zhuo; Eddy, Douglas; Krishnamurty, Sundar.
"Self-Improving Additive Manufacturing Knowledge Management."
Paper presented at Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, Quebec City, Canada. August 26, 2018 - August 28, 2018.

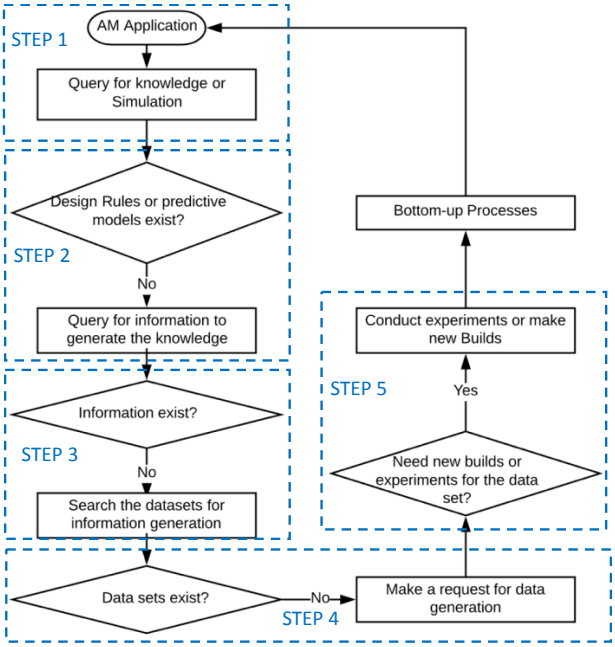


Figure 7: A typical top-down process

By far, few research activities on streamlining goal-oriented AM data generation process exist. In information fusion for combat system command and control, a user directed information discovery method has been proposed to manage intelligence products for mission objectives [12]. A similar methodology can be developed for AM information systems.

3.3 Closed-loop Knowledge Management

As shown in both Figure 6 and Figure 7, the bottom-up process and the top-down process are naturally connected with each other at the end, which forms a closed loop AM knowledge lifecycle. The closed-loop system, consisting of the bottom-up adaptive knowledge engineering and the top-down data generation process, if automated, will not only accelerate AM knowledge acquisition, and also reduce AM data generation cost dramatically. Data and information sharing is critical to implement an automated self-improving knowledge management system. Collaborative AM development platforms built on a shared knowledge base will further shorten the AM product, machine, material and process development lifecycle. Service-oriented architecture has the potential to integrate all the tiers and the links into an organic eco-system. For those links involving human-in-the-middle, a publish/subscribe mechanism can improve human response to knowledge changes and new data generation requests.

4 A CASE STUDY: A METAL AM PROCESS MODEL ADAPTATION

A case study demonstrates a manual process for bottom-up adaptive AM knowledge management. A predictive metamodel was adapted for this purpose and deployed with empirical data from a metal AM process. This approach uniquely combines a previously established metamodel of the process with newly ingested data to complete a self-improving process. The general workflow deployed here is shown in Figure 8.

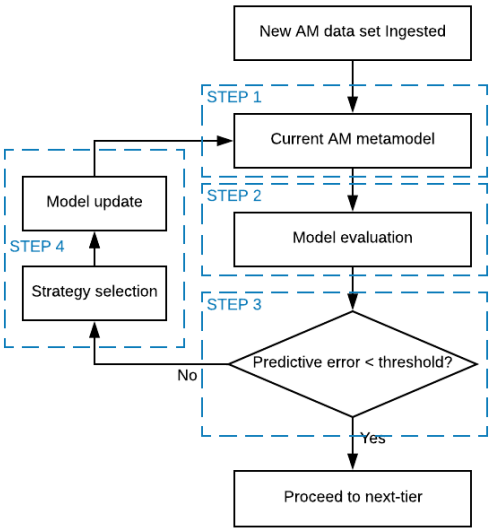


Figure 8. General workflow of data-driven metamodel updating

As shown in Figure 8, newly ingested data triggers an evaluation of the current metamodel for prediction accuracy. The prediction error of the metamodel on the new data set is used for decision making for model update. If the prediction error is within a predefined threshold, the metamodel is considered effective, and no modification is needed. Otherwise, the model has to be improved using the new data and with the best available update strategy. In this case study, the predefined threshold value is generated from the average relative error magnitude (AREM) of the current metamodel based on the leave-one-out (LOO) cross-validation method [21, 22]. Three update strategies were designed for model updating, including 1) Direct data combination – assuming that the existing metamodel was trained and validated using the data generated under the same experimental condition as the new one, new data points are just added to optimize the model parameters. 2) Grey-box modeling [11], which is applied if the new experimental conditions are different, but the design spaces are highly overlapped. 3) Interpolation modeling strategy, which can be chosen for any cases. Here again, AREM [10][21] is selected to evaluate the predictive accuracy:

$$AREM = \frac{1}{n}\left(\frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{y_i}\right)$$

where y is the observed value, $\hat{y}$ is the value predicted by the metamodel and n is the total number of new data points.

Lu, Yan; Yang, Zhuo; Eddy, Douglas; Krishnamurty, Sundar.
"Self-Improving Additive Manufacturing Knowledge Management."
Paper presented at Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, Quebec City, Canada. August 26, 2018 - August 28, 2018.

For our study, data sets were collected from two independent experiments conducted for a laser melting powder bed fusion metal AM process [23]. The first experiment melted the powder directly on a bare build plate. This set of data is used to build an initial metamodel. The second experiment was conducted on a single layer of powder and the data generated is considered as a new data set. Laser power (LP) and scanning speed (SS) are the input variables, and melt-pool width is the output variable of the metamodel. Both experiments used the fractional factorial design of experiments (DOE) method, with the laser power ranging from 100W to 250W and the scan speed ranging from 200 μm/s to 1400 μm/s. Two data points were left out because of infeasible builds at low energy intensity. The remaining 26 data points are listed in Table 2. The melt-pool width was measured along a 1mm section near the center of the scan trace and the measurement method is detailed in Fox et al [25]. The mean value is listed in the table.

Table 2. Results of laser melting experiments

LP (W)	SS (mm/s)	Melt-pool width (μm)	
		Bare build plate	On powder
100	200	134.57	127.77
100	400	114.75	112.43
100	800	80.52	97.98
100	600	87.58	86.64
100	1000	75.35	64.43
150	200	181.44	162.65
150	400	126.50	149.24
150	600	124.70	129.07
150	800	106.39	119.95
150	1000	103.50	101.26
150	1200	99.28	97.98
150	1400	99.40	95.95
195	200	235.94	225.16
195	400	178.07	150.01
195	600	150.52	153.05
195	800	129.57	151.04
195	1000	122.86	119.19
195	1200	115.38	125.60
195	1400	112.40	114.83
250	200	247.39	253.57
250	400	227.55	254.10
250	600	159.31	150.13
250	800	160.85	175.71
250	1000	141.34	141.05
250	1200	134.58	137.31
250	1400	126.69	124.42

To generate an appropriate metamodel update strategy, we manipulated and divided the original data points into several sub data sets to represent various data-model matching scenarios. Table 3 summarizes four sub data set partitions covering four possible data-model matching scenarios. For scenarios 1 and 2, both the initial data set and the new data set are sampled from Experiment 1. For scenario 1, 15 out of 26 data points from the first experiment were sampled by space filling to construct the initial metamodel. The other 11 points are treated as newly ingested data. In scenario 2, the design space of the initial metamodel and the domain of the new data set are manipulated to be only partially overlapped. Scenarios 3 and 4 assume data from the first experiment for the initial metamodel construction and select data from the second experiment to form a new data set. Scenarios 3 and 4 consider different design space overlapping conditions.

Table 3. Test problem scenarios

No.	Overlapping condition				Experimental consistency
	Initial		New		
	LP	SS	LP	SS	
1	100-250	200-1400	100-250	200-1400	Yes
2	100-150	800-1400	195-250	200-1000	Yes
3	100-250	200-1400	100-250	200-1400	No
4	195-250	200-1400	100-195	200-1400	No

Polynomial regression was selected to build a metamodel in this example considering the linearity observed between the inputs and the output. If a grey-box modeling strategy is applied, Kriging method [24] is combined with the initial polynomial model. Table 4 lists the model update results for all test scenarios. The initial AREM was measured using the LOO cross-validation method. Predictive AREM refers to the prediction error on the new data points. Final AREM, however, is calculated differently for different scenarios. Scenarios 1, 2 and 4 use the LOO cross-validation method since there is no third data set available within the given design space. Scenario 3 used a third data set to evaluate the updated AREM beyond the initial LOO cross-validation. The third data set is extracted from the second experiment result, which covers those points not selected to construct the newly ingested data (with results shown in parenthesis). A star symbol marks the best update strategy and corresponding AREM.

Table 4. Results from model updating after ingesting new data

No.	Initial AREM	Predictive AREM	Updating strategy	Final AREM
1	0.0783	0.0626	Direct combination	0.0528
2	0.0520	0.3614	Direct combination	0.0760
			Interpolation modeling*	0.0743*
3	0.0528	0.0700	Direct combination	0.0564 (0.0749)
			Grey-box modeling*	0.0001* (0.0670*)
			Interpolation modeling	0.0822 (0.0852)

Lu, Yan; Yang, Zhuo; Eddy, Douglas; Krishnamurty, Sundar.
"Self-Improving Additive Manufacturing Knowledge Management."
Paper presented at Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, Quebec City, Canada. August 26, 2018 - August 28, 2018.

4	0.0672	0.2074	Direct combination*	0.0585*
			Interpolation modeling	0.0598

For Scenario 1, the predictive AREM on the new data set is 0.0626 which is less than the initial AREM. This observation indicates that it might be unnecessary to update the current metamodel further since the model can accurately predict the new data points. However, it is observed that after the model updating using the direct data combination strategy, the final AREM is further reduced to 0.0528. Scenario 2, however, requires a major update for the initial metamodel since it generates very large predictive AREM (0.3614) on the new data set. Both applicable strategies provided model improvement at a similar level. Though the AREM of the direct combination method is slightly higher than that of the interpolation strategy, it fundamentally increases the design space. For scenario 3, 13 data points are picked from the second experiment to form the new data set. The rest of the data is available to validate the model since the two data sets are in the same design space. In this case, the interpolation strategy is not optimal as the updated model generated has higher AREM. Instead, grey-box updating results in the greatest improvement. Scenario 4 represents the most complex situation among these test scenarios. Both design space and experimental conditions are inconsistent between the initial model and the new data set. Thus, any strategy that can utilize more information to update the metamodel would become useful. After strategic updating, the AREM was reduced from 0.2074 to 0.0585.

The case study demonstrates a valid approach to updating an existing metamodel using new data. It sheds a little light on a general metamodel update approach. However, the case study only illustrates a manual process for model updating. Further research is needed to discover an automated and more effective way for metamodel updating following emergence of new data sets in the AM domain.

5. CONCLUSIONS

As AM matures into a production-ready technology, greater emphasis will continue to be placed on rapid design-to-product transformations. AM will continue to become a more viable alternative for applications such as supply chain logistics and customized parts. To this end, this paper outlined a closed-loop data-information-knowledge-application framework that will support the functionalities necessary to realize rapid, customizable, design-to-product transformations through a self-improving knowledge management system.

The proposed analytic framework defines a bottom-up knowledge engineering process and a top-down data generation process to leverage individual efforts of conducting experiments and deriving knowledge from data. The streamlined bottom-up knowledge engineering process plus the application driven data generation process are connected by operators and engineers. As new data is ingested, it will be infused to the existing information system. The contextualized new data could trigger a metamodeling process where predictive models are updated to reflect PSP more accurately. Sequentially, the new knowledge will be integrated into the AM application to improve AM engineering decisions. Conversely, if the engineer receives some alarm caused by inconsistency between design decisions and design rules, he/she can perform engineering analysis by querying the knowledge. If knowledge is missing, information needs will be generated, and the demand will be passed to the information system, and ultimately design-of-experiments can be prescribed for the researchers and manufacturers. Thus additional experiments, builds, and tests can be performed in a cost effective fashion.

The integrated and automated workflow illustrated in this paper is comprehensive enough to cover and manage the whole development lifecycle of AM knowledge with continuous improvement. It was also discovered that to automate and streamline the workflow, further research on AM informatics is needed to build the links still missing, especially those in the top-down goal-oriented data generation processes.

Our future work will be focused on the integration of knowledge-based adaptive data-driven knowledge engineering and goal-oriented adaptive data sampling design. Specifically, we will investigate how ontology can potentially address the aforementioned need to learn from the experience of AM parts produced successfully or unsuccessfully. This would reduce requirements to simulate or analyze each new job fully. Approaches will be proposed to adaptively learn and enrich the knowledge base to enable continuous improvements. The two-way concept introduced in Section 3 provides the foundation to methodically adapt data and knowledge to improve the quality, reliability, and usefulness of both for improved understanding of the complexity of PSP relationships. This approach can help move toward a reusable knowledge base that improves with experience. A reference ontology can be developed and standardized to enable easy integrations of heterogeneous AM information systems.

DISCLAIMER AND ACKNOWLEDGEMENT

Certain commercial equipment, instruments, or materials identified in this paper are not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

The authors would like to thank Dr. Jason Fox for the experimental data used in our case study. We also feel thankful for the support and the insightful discussions provided by Dr. Paul Witherell and Dr. Alkan Donmez at NIST.

REFERENCES

[1] Cotteleer, M., “3D opportunity for production: Additive manufacturing makes its (business) case”. Deloitte Review, Issue 15, 2014

Lu, Yan; Yang, Zhuo; Eddy, Douglas; Krishnamurty, Sundar.
”Self-Improving Additive Manufacturing Knowledge Management.”
Paper presented at Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, Quebec City, Canada. August 26, 2018 - August 28, 2018.

[2] The Additive Manufacturing Software Conundrum, http://www.padtinc.com/blog/the-focus/the-additive-manufacturing-software-conundrum, Accessed on March 10, 2018

[3] https://www.3dhubs.com, Accessed on March 10, 2018

[4] Optimize your Additive Manufacturing Know-How, http://www.digitaleng.news/de/optimize-your-additive-manufacturing-know-how/, Accessed on March 10, 2018

[5] Albright, B., Design for 3D Printing, http://www.digitaleng.news/de/design-3d-printing/, Accessed on March 10, 2018

[6] Lu, Y., Witherell, P. and Lopez, F., Digital Solution for Additive Manufacturing, ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Charlotte, NC, August 2016

[7] Kim, D.B., Witherell, P., Lipman, R., Feng, S.C., "Streamlining the additive manufacturing digital spectrum: A systems approach". Additive Manufacturing, Issue 5, 2015, pp. 20-30

[8] Lu, Y., Choi, S., Witherell, P., "Towards an integrated data schema for additive manufacturing conceptual modeling". ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Boston, MA, August 2015

[9] Lu, Y., Witherell, P and Donmez, A., A Collaborative Data Management System for Additive Manufacturing, ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Cleveland, OH, August 2017

[10] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., Lu, Y. and Witherell, P., Investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing. In ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Cleveland, OH, August 2017

[11] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., and Lopez, F., Investigating predictive metamodeling for additive manufacturing. In ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Cleveland, OH, August 2016

[12] Kessler, O. and White, F., "Data Fusion Perspectives and Its Role in Information Processing," Ch.2 in Handbook of Multi-Sensor Data Fusion 2nd Ed, (eds.). Liggins, M. E., Hall, D., and linas, J. L, CRC Press, 2008.

[13] Rowley, J. "The wisdom hierarchy: representations of the DIKW hierarchy". Journal of Information and Communication Science. **33** (2): 163–180, 2007

[14] Mathematical Markup Language, https://www.w3.org/TR/WD-math-980106/, Accessed on March 10, 2018

[15] Data Mining Group, "PMML 4.0 - Mining Schema". http://dmg.org/pmml/v4-0-1/MiningSchema.html, Accessed on March 10, 2018

[16] Eddy, D., Krishnamurty, S., Grosse, I., Perham, M., Wileden, J., & Ameri, F., 2015, "Knowledge Management with an Intelligent Tool for Additive Manufacturing." ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Volume 1A: 35th Computers and Information in Engineering Conference, Boston, MA, August 2015

[17] Hagedorn, T., Krishnamurty, S., Grosse, I., " A Knowledge-Based Method for Innovative Design for Additive Manufacturing Supported by Modular Ontologies," Journal of Computing and Information Science in Engineering, JCISE-17-1272, accepted manuscript

[18] Ekaputra, E., Serral M., Kiesling, E., and Biffl, S., Ontology-Based Data Integration in Multi-Disciplinary Engineering Environments: A Review, Open Journal of Information Systems (OJIS) Volume 4, Issue 1, 2017

[19] Jee, H., Lu, Y., and Witherell, P, 2015. "Design rules with modularity for additive manufacturing". International Solid Freeform Fabrication Symposium, Austin, TX, pp. 1450-1462.

[20] Lopez, F., Witherell, P., and Lane, B., "Identifying uncertainty in laser powder bed fusion models", 2016 ASME Manufacturing Science and Engineering Conference, Blacksburg, VA, 2016

[21] Refaeilzadeh P, Tang L, Liu H. Cross-validation. In Encyclopedia of database systems (pp. 532-538). Springer US, 2009

[22] Shao T, Krishnamurty S. A clustering-based surrogate model updating approach to simulation-based engineering design. Journal of Mechanical Design. 2008 Apr 1;130(4):041101.

[23] Fox, J.C., Whiting, J., Yeung, H., Lane, B., Gratham, S., Neira, J., Fisher, B., "Initial Scan Tests of the NIST Additive Manufacturing Metrology Testbed (AMMT)," Presented at the 2016 SFF Symposium, Austin, TX. August 8-10, 2016

[24] Cressie, N., Statistics for spatial data. John Wiley & Sons, 2015

Lu, Yan; Yang, Zhuo; Eddy, Douglas; Krishnamurty, Sundar.
"Self-Improving Additive Manufacturing Knowledge Management."
Paper presented at Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, Quebec City, Canada. August 26, 2018 - August 28, 2018.

[25] Fox, JC., Lane, BM, Yeung, H, "Measurement of process dynamics through coaxially aligned high speed near-infrared imaging in laser powder bed fusion additive manufacturing", Proc. SPIE 10214, Thermosense: Thermal Infrared Applications XXXIX, 1021407 (5 May 2017); doi: 10.1117/12.2263863

Lu, Yan; Yang, Zhuo; Eddy, Douglas; Krishnamurty, Sundar.
"Self-Improving Additive Manufacturing Knowledge Management."
Paper presented at Proceedings of the ASME 2018 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, Quebec City, Canada. August 26, 2018 - August 28, 2018.

StormSense: A Blueprint for Coastal Flood Forecast Information & Automated Alert Messaging Systems

Jon Derek Loftis
Virginia Institute of Marine Science,
College of William & Mary
Gloucester Point, VA, U.S.A.
e-mail: jdloftis@vims.edu

Sridhar Katragadda
City of Virginia Beach,
Dept. of Comm. and Info. Technology
Virginia Beach, VA, U.S.A.
e-mail: SKatraga@vbgov.com

Sokwoo Rhee and Cuong Nguyen
Smart Grid & Cyber-Physical Systems Program Office,
National Institute of Standards and Technology
Gaithersburg, MD, U.S.A.
e-mails: sokwoo.rhee@nist.gov
& cuong.nguyen@nist.gov

***Abstract*— Increased availability of low-cost water level sensors communicating through the Internet of Things (IoT) has expanded the horizons of publicly-ingestible data streams available to modern smart cities. StormSense is an IoT-enabled inundation forecasting research initiative and an active participant in the Global City Teams Challenge seeking to enhance flood preparedness in the smart cities of Hampton Roads, VA for flooding resulting from storm surge, rain, and tides. In this study, we present the a blueprint and series of applicable protocols through the use of the new StormSense water level sensors to help establish a regional resilience monitoring network. In furtherance of this effort, the Virginia Commonwealth Center for Recurrent Flooding Resiliency's Tidewatch tidal forecast system is being used as a starting point to integrate the extant (NOAA) and new (USGS and StormSense) water level sensors throughout the region, and demonstrate replicability of the solution across the cities of Newport News, Norfolk, and Virginia Beach within Hampton Roads, VA. StormSense's network employs a mix of ultrasonic sonar and radar remote sensing technologies to record water levels and develop autonomous alert messaging systems through the use of three separate cloud environments. One to manage the water level monitoring sensors and alert messaging, one to run the model and interface with the post-processed results, and one to geospatially present the flood results.**

***Keywords*—Hydrodynamic Modeling, Internet of Things, Smart City, Global City Teams Challenge, Replicability, Citizen Science, Sea Level Rise**

I. Introduction

Modern smart cities are functionally equivalent to a complex system. These systems are often subjected to many non-linear influences on how to efficiently allocate their limited resources [1]. The protocols used to determine how smart cities respond to emergency flooding conditions in the future could be adapted using models. These models should be informed and validated by a dense water level sensor network to most efficiently advise how best to prepare for the imminent flood-related disasters of the future [2,3]. As many data-driven projects are afforded greater versatility on cloud based platforms, StormSense approaches flood monitoring and modeling via the cloud. This is done through the use of IoT-sensor monitoring, online mapping, and through predictive tidal and hydrodynamic modeling with automated alerts [4].

StormSense is a flood prediction project initiated by the proactive local governments in tidewater Virginia. It detects, models, and communicates flood risk with help from scientists at the Virginia Institute of Marine Science (VIMS) and partner city engineers via IoT water level sensors, hydrodynamic models, artificial intelligence, and voice-assisted technologies. StormSense operates and disseminates flood forecasts via a

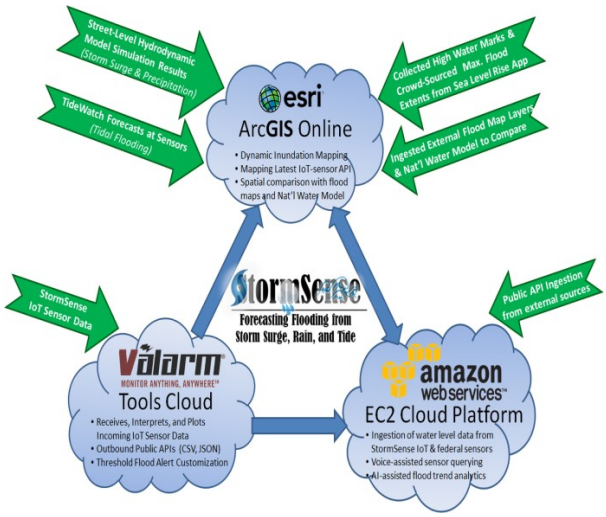


Figure 1. StormSense's triumvirate of cloud platforms employed for automated flood alert messaging. Inputs noted in green arrows, with blue arrows depicting exchange between cloud platforms. Click figure for larger view.

Nguyen, Cuong; Rhee, Sokwoo; Loftis, Jon 'Derek'.
"StormSense: A Blueprint for Coastal Flood Forecast Information & Automated Alert Messaging Systems."
Paper presented at Third International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership with Global City Teams Challenge (GCTC), Porto, Portugal. April 10, 2018 - April 10, 2018.

triumvirate of cloud platforms operated by Valarm, ESRI, and Amazon Web Services (AWS) (Figure 1):

1) Valarm's IoT water level sensors densify the existing model data matrix and help us better understand the varying wind conditions that cause recurrent ephemeral flooding.

2) AWS' cloud platform aids with smart voice-assisted technologies using Reverb/Amazon Alexa, to place flood observations and predictions in citizens' hands visibly and audibly via their smart devices and StormSense's AWS.

3) ESRI's ArcGIS Online mapping environment visually disseminates flood model forecasts. Citizens find that flood layers overlapping their house, driveway, or route to work, is difficult to misinterpret.

The computationally-intensive nature of the hydrodynamic modeling approach is such that StormSense currently operates with the limitation that storm surge and heavy rainfall forecasts require the model to be manually submitted for simulation via high performance computing platfo rms or on AWS' EC2 cluster [4]. However, tidal flooding forecasts are automated through a service VIMS operates called Tidewatch. Hydrodynamic modeling of storm surge requires the implementation of a large-scale regional model to accurately capture the large scale wind influence of hurricanes and nor'easters as their storm surges transition from the open ocean to Atlantic Shelf, into Chesapeake Bay and then into its contributing estuaries. Heavy rainfall events are complex to model, partially because precipitation observation data for the region are currently logged in an aging system architecture by semi-private regional entities, and also due to lack available

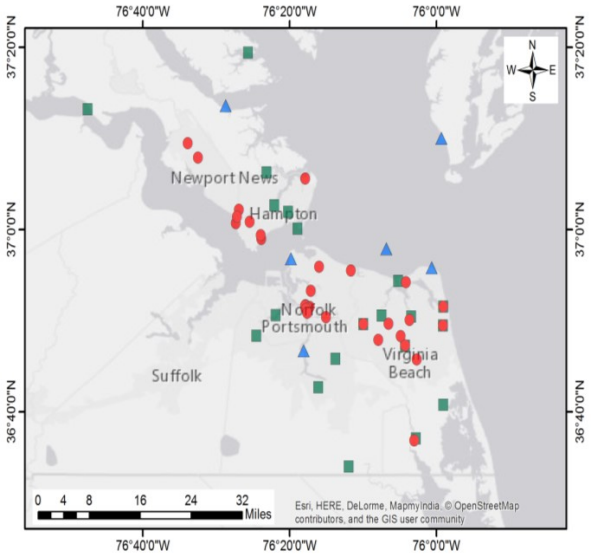


Figure 2. Map of 57 publicly-streaming water level monitoring stations throughout Hampton Roads, VA. The StormSense sensor network has contributed 28 sensors to the 29 existing sensors maintained by federal entities. Of these, NOAA has 6 (marked in blue) and USGS maintains 19 (noted in green). Additionally, VIMS has 1, and WeatherFlow has 3 (also marked in red). Click figure or http://arcg.is/14aCe1 for interactive map.

higher-order temporal and spatial resolution data to most accurately forecast heavy precipitation events [5].

Despite these limitations, the most frequent form of flooding experienced by the localities in Hampton Roads is tidal nuisance flooding [6]. Alert messaging for tidal flooding events can be addressed through a completely automated approach which can make use of sensors and harmonic tidal signature extraction techniques. These methods can be harnessed to estimate when tidal flooding is likely to occur at or near a sensor, and automate alerts associated with designated flood thresholds as an automated and advanced early warning system desired by coastal smart cities to protect citizens, infrastructural assets, and qualify for decreases in flood insurance premiums proportional to the alert system's advance warning time and sophistication. As Tidewatch currently provides tide forecasts up to 36 hours in advance of tidal inundation events, this approach is desirable for smart communities participating in FEMA's National Flood Insurance Program.

Thus, the StormSense Project brings together municipal governments in Hampton Roads, Virginia, including: Newport News, the RSCT grant recipient, Norfolk, Virginia Beach, Hampton, Chesapeake, Portsmouth, Williamsburg, and York County along with VIMS, to develop a regional resilience monitoring network [6]. This network of 28 newly-installed publicly-broadcasting water level sensors ingests and interfaces with open Application Programing Interface (API) data from federal monitoring and water prediction agencies (such as USGS and NOAA) to bring the total number of water level sensors to 57 (Figure 2). With most of these sensors being recently installed in 2016-2017, StormSense is poised to develop the network as Phase 1 [7], and develop a street-level flood forecasting and monitoring solution across the entire region for Phase 2 [4], which begins with integration of observed water-levels into VIMS' Tidewatch tidal forecasting system, which now operates under the Virginia Commonwealth Center for Recurrent Flooding Resiliency (CCRFR) at: https://www.floodingresiliency.org/ [3].

Many existing smart cities solutions are designed to have a measurable impact on specific key performance indicators relevant to their communities. Since many of today's smart city/community development efforts are often isolated and highly customized projects, the National Institute of Standards and Technology (NIST) launched the Global City Teams Challenge (GCTC) to encourage collaboration and the development of standards for smart cities. The GCTC's long-term goal is to demonstrate a scalable and replicable model for incubating and deploying interoperable, adaptable, and configurable Internet of Things (IoT)/Cyber-Physical Systems technologies in smart cities/communities [1]. This program aims to help communities benefit from working with others to improve efficiency and lower costs. NIST also created the Replicable Smart City Technology (RSCT) cooperative agreement program to provide funding to enable awardee City/Community Partners to play a lead role in the team-based

Nguyen, Cuong; Rhee, Sokwoo; Loftis, Jon 'Derek'.
"StormSense: A Blueprint for Coastal Flood Forecast Information & Automated Alert Messaging Systems."
Paper presented at Third International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership with Global City Teams Challenge (GCTC), Porto, Portugal. April 10, 2018 - April 10, 2018.

GCTC effort to pursue measurement science for replicable solutions [8]. The RSCT program was designed to support standards-based platform approaches to smart cities technologies that can provide measurable performance metrics. Together these two programs work to advance state-of-the-art smart city standards. The city of Newport News was awarded an RSCT grant in September 2016 on behalf of the StormSense Project's application to NIST. Thus, the implementation of open-source models, accessible cloud platforms, and low-cost IoT solutions ideally embody the GCTC mantra by make StormSense's solution tenably replicable, scalable, and measurable [7]. The combined nature of this approach ideally will not only make a difference in our region, but potentially in other flood-prone regions of the world through the use of the blueprint presented in the next section and the data ingestion protocols noted throughout this paper.

II. Study Area and Model Blueprint

Hampton Roads, VA, is the second-largest population center at risk from sea level rise in the United States. The region has more than 400,000 properties that are exposed to flood or storm surge inundation [9]. The region has a population of over 1.7 million people, living and traveling on roads exposed to both severe and increasing frequent chronic "nuisance" flooding [10,11]. A major issue Hampton Roads faces is that the region experiences nuisance flooding fatigue with such frequency that it is easy to forget that flooding events cost our cities, their first responders, and their residents time and money [12]. In one neighborhood in the City of Newport News particularly prone to nuisance flooding, typically many emergency responders were required to assist in evacuating the complex [13,14]. However, by remotely alerting residents that the water was rising quickly on the local stream, the past two flooding events have not required any emergency responders to assist in evacuating and were subsequently able to dedicate their emergency services elsewhere [14,15]. The goal of establishing a flood monitoring network can be expensive, but in the long term, the anticipated benefits of improved quality of life for a region's citizens are monumental. The goal is to replicate this level of success throughout the cities of Hampton Roads by providing a greater density of water level sensors. As an added benefit, more publicly-available water level sensors empower property owners to take responsibility for their assumed risk of living adjacent to floodplains. This has resulted in a marked spike in the number of residents who have opted for flood insurance, with 2,231 claims totaling $25M in damage attributed to 2016 Hurricane Matthew [5,16]. Many of these properties are insured through the Federal Emergency Management Agency's (FEMA) National Flood Insurance Program, but many properties outside of the designated floodplain do not have preferred risk policies [12,16]. Thus, StormSense has developed a blueprint which has been shared with the GCTC community via the Public Safety Supercluster at the 2017 GCTC Expo in Washington, D.C. (Figure 3)[17].

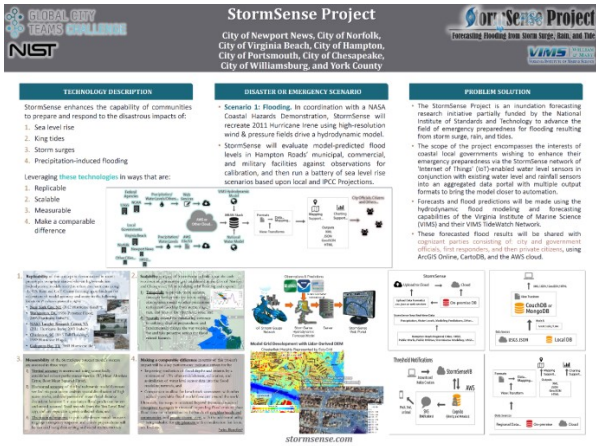


Figure 3. GCTC Blueprint Solution for StormSense flood monitoring, model predictions, and automated alert system protocols. Click figure for larger view.

Existing flood communication and messaging systems have not yet responded to the changing risk patterns brought by sea level rise and have not been able to meet the diverse needs of a growing populous in an expanding floodplain. Thus, a better understanding of flood risk perception, information seeking behavior and decision-making can inform the development of new communications tools and flood risk messaging [18]. This is the percieved intersect between new IoT-technologies and emerging flood model validation methods. For each storm event, in Hampton Roads, water levels driven via 36-hour Tidewatch forecasts provided by VIMS at NOAA's Sewells Point gague are typically used to drive surge and tides in urban-scale models. Now, forecasts from any of Tidewatch's ingested data from StormSense sensors can be used as the model's boundary conditions alongside wind and pressure inputs used to drive the model atmospherically, similar to Loftis, Wang, and Forrest [19]. VIMS employs a street-level hydrodynamic model, which incorporates a non-linear solver and variable sub-grid resolutions [2], capable of being embedded with lidar-derived topography to scale resolution for inundation where it is needed down to 5-m or even 1-m resolution in known areas where flooding frequency is high [19,20]. The model has been used to simulate every major storm event in Hampton Roads that has occurred in the past 25 years, and has been used in many other places along the U.S. East and Gulf Coasts as well [5,21-24]. For more information on the hydrodynamic models, please refer to these cited studies.

III. Water Level Sensors

StormSense has recently deployed 28 bridge-mounted IoT-ultrasonic and microwave radar water level sensors in Newport News, Virginia Beach, and Norfolk, as outlined on the StormSense project's website at: http://www.stormsense.com. These sensors will complement the previously installed array of 6 gauges operated by NOAA, 19 relatively new gauges

Nguyen, Cuong; Rhee, Sokwoo; Loftis, Jon 'Derek'.
"StormSense: A Blueprint for Coastal Flood Forecast Information & Automated Alert Messaging Systems."
Paper presented at Third International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership with Global City Teams Challenge (GCTC), Porto, Portugal. April 10, 2018 - April 10, 2018.

installed in 2015-2016 via Hurricane Sandy relief funds operated by the USGS, and 1 gauge operated by VIMS in Hampton Roads (Figure 2). While the extant remote sensors in the region are largely K_a-band radar sensors transmitting data through satellite signals, the new StormSense IoT-sensors enlist the use of ultrasonic sensors and transmit data via cellular transmission protocols or Long Range (LoRa) Wireless Area Networks (WAN), with the focus of creating a replicable cost-effective network of sensors. Some realized utilities for a dense network of water level sensors are noted as follows:

1) Archiving water level observations for flood reporting
2) Validation/inputs for hydrodynamic flood models
3) Automated targeted advance flood alert messaging
4) Reliable interpolation of continuous water surface elevations throughout geo-event processing capabilities

A. Water Level Sensor Types and Applications

A collaboration between VIMS and the partner cities of: Newport News, Hampton, Norfolk, Virginia Beach, Portsmouth, Chesapeake, Williamsburg, and York County, in Hampton Roads, VA, will provide a prototype for strengthening emergency response times by providing spatial flood extent predictions in interactive map form at 5-m resolution. The plan for integrating the inundation model into a more permanent warning system involves planned connection with the new sensors to the cities' current Everbridge notification systems for alert messaging. This occurs when the sensor observes flooding at user-specified elevations, and integration with model predictions for timely forecasted tidal inundation alerts through Tidewatch once the sensors are tidally-calibrated. In Newport News, the city employed a mix of 2 K_a-band radar sensors and 6 ultrasonic sonar sensors from Valarm, a California-based sensor vendor with a cloud-based virtual alarm messaging platform. The Valarm Tools Cloud platform uses the newly-installed sensors to provide subscriber-based alerts (Figure 4) based upon water level observations. The system will also eventually ingest tidal forecast predictions once incorporated into Tidewatch to enable cities to provide a unique flood-preparedness service to their citizens. An added benefit to this automated flood alert messaging method is that it can bolster the flood warning portion of their FEMA NFIP application to participate in the Community Rating System (CRS). This is important, as each higher participation level the city achieves in the hierarchical CRS program is commensurate with an additional 5% decrease in flood insurance premiums for the citizen homeowners in participating communities.

StormSense demonstrates the benefits of replicating shared smart city solutions across multiple cities and communities that are facing similar flood challenges and it aligns with the goals of GCTC and RSCT programs [8]. For a different innovative example, Norfolk's LoRaWAN ultrasonic sensor network was established in city's historic Hague region in August 2017. The sensor network is currently comprised of one tide monitoring sensor mounted over The Hague walking bridge near where the USGS mounts their temporary rapid deployment gauge, and five inundation sensors, strategically positioned over frequently flooded streets [4,6]. The LoRaWAN sensors were purchased through a Norfolk-based vendor, GreenStream, Inc., and use long range WiFi instead of cellular data transmissions. They are currently publicly reporting water level observations in Tidewatch, and public API URLs are available at: http://www.vims.edu/people/loftis_jd/HRVASensorAssets/index.php.

The recent installation of water level sensors provided by the USGS were used as an opportunity to demonstrate some of the benefits of added water level sensors using these ultrasonic sensors will be evaluated as reputable and replicable monitoring methods after a longer-term study. In pursuit of this, Figure 5 shows three examples of temporary StormSense ultrasonic sensors deployed on the same bridges as the USGS' K_a-band radar sensors over tidal rivers and creeks throughout the City of Virginia Beach [4]. A later paper will evaluate the differences between these sensor accuracies and types, fault tolerance in data transmissions, and solar power management schemes [6]. An initial comparison with a temporary rapid deployment gauge established by the USGS allowed for a

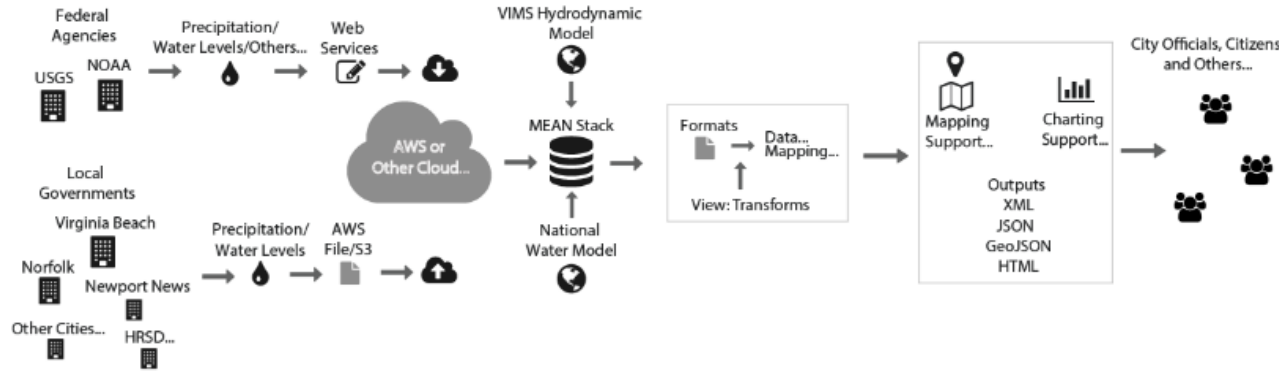


Figure 4. StormSense's information flow to guide Hampton Roads' data ingestion efforts to advise predictive flood models.

Nguyen, Cuong; Rhee, Sokwoo; Loftis, Jon 'Derek'.
"StormSense: A Blueprint for Coastal Flood Forecast Information & Automated Alert Messaging Systems."
Paper presented at Third International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership with Global City Teams Challenge (GCTC), Porto, Portugal. April 10, 2018 - April 10, 2018.

favorable short-term data comparison with Norfolk's LoRaWAN sensor collocated there during a nine-day overlap period during Hurricane Maria [4].

B. Water Level Sensor Accuracies and Costs

After an evaluation period of 6-9 months, these collocated StormSense sensors will be relocated to unique monitoring locations in Virginia Beach. A small number of white papers and vendor brochures evaluate the accuracies of the ultrasonic and K_a-band radar sensors in laboratories or for the application of level monitoring of water treatment reservoirs or chemical vats. However, these are not comparable to tidal water bodies or areas with significant wave action, such as during the extratropical storm surge events presented in this study during Hurricanes Jose and Maria [6]. A cursory comparison from the initial deployments of the sensors in Summer 2017 revealed that the ultrasonic sonar units are from Valarm are accurate in the lab to a Root Mean Squared Error (RMSE) of ±5 mm, and accurate in the field to an average of ±18 mm, while the two K_a-band radar sensors in Newport News are accurate in the lab to ±1 mm and accurate as deployed in the field to ±9 mm [7]. The cost to purchase a solar-powered cellular transmission station was approximately $3000/each for the ultrasonic sensors, and $4400/each to purchase the K_a-band radar units [4]. The street inundation sensors employed in Norfolk through the vendor, Green Stream, are accurate in the lab to approximately ±15 mm, and accurate in the field ±45 mm, and sensors were purchased for $400/each, plus the cost of the LoRa transmission gateway, which has an effective transmission range of approximately one mile, less the distances occluded by high-rise buildings [7].

C. Water Level Sensor Data Comparisons

A comparison of the five new street inundation sensors and one water level sensor in Norfolk, and eight new water level sensors in Newport News were used to temporally and vertically validate a street-level hydrodynamic model's predictions during the offshore passage of Hurricanes Jose and Maria, which detected increased water levels in Hampton Roads by 76.2 cm (2.5 ft.) and 60.9 cm (2 ft.), respectively. These six gauges resulted in an aggregate vertical RMSE of ±8.93 cm over a 72-hour Hurricane Jose model forecast simulation [4].

The seven gauges present during Hurricane Maria (including the USGS rapid deployment gauge installed from 9/21-9/29/2017) yielded a more favorable aggregate RMSE of ±6.28 cm when compared with the model. Both storms produced minimal surge-related coastal flooding, yet inundation impacts were equally profound in some tidal-connected inland areas, making the comparison with Norfolk's new street inundation sensors interesting to observe and practical for verification of inland inundation extents and depths. USGS measurements temporarily co-located at the same site during Hurricane Maria's passage were used to apply a vertical adjustment of +4.5 cm (0.15 ft.), based upon the Mean Absolute Error (MAE) as an offset, to improve the Root Mean Squared Error (RMSE) metric for this event, and likely for many events in the future [4]. This change resulted in an improvement in sensor estimated RMSE from 6.08 to 0.71 cm, a difference of 5.37 cm (0.17 ft.).

IV. Discussion and Conclusions

In the future, smart city systems could evaluate tenable candidate blueprint solutions for flood-related problems, whether they be attributed to storm surge, heavy rainfall, or tides, as was the case during the offshore passage of 2017 Hurricanes Jose and Maria, using a decision matrix. This could help key decision-makers in areas analogous to Hampton Roads, VA, make informed decisions regarding how flood-related solutions could be best addressed in their region. The new StormSense water level sensor network is being integrated into Tidewatch to create a regional resilience monitoring network to directly address a key recommendation from the area's Intergovernmental Pilot Project [3,25].

StormSense's value can be inherently measured in: time, money, and potential lives saved. For evidence of this, the project and the region's media partners asked citizens to put the new water level sensors and flood model predictions to the test during the 2017 king tide floods on 5 November, 2017. 500+ volunteers from 12 cities and counties helped to map 53,000+ GPS-recorded high water marks and collected 1,200+ geotagged photographs of flooding in Hampton Roads using a mobile app called Sea Level Rise [26]. Overall, the event revealed that the new StormSense sensors and model were vertically accurate within a root mean squared error of 2.21 cm and horizontally accurate within a mean spatial distance difference of 13.1 ft. [6,26].

Figure 5. Examples from three StormSense ultrasonic sonar sensors co-located in the field adjacent to USGS K_a-Band radar sensors in Virginia Beach for direct comparison of monitoring accuracy. These sensors will temporarily be stationed adjacent to each other for a period of 6-9 months to provide a long term data record for comparison of water level measurements, data transmission speeds, and solar power efficiency. Figure adapted from [4].

Nguyen, Cuong; Rhee, Sokwoo; Loftis, Jon 'Derek'.
"StormSense: A Blueprint for Coastal Flood Forecast Information & Automated Alert Messaging Systems."
Paper presented at Third International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership with Global City Teams Challenge (GCTC), Porto, Portugal. April 10, 2018 - April 10, 2018.

As sea levels rise, it is likely that this will continue to become an ever more pervasive issue. Analysis of the local sea level trend from the longest period record in Hampton Roads at Sewells Point in the City of Norfolk depicts a long-term linear increase in mean sea level of 4.59±0.23 mm/year since its establishment in 1928 [27]. The data from a new sea level trend study conducted at VIMS focuses on trends since the Anthropocene (1969-present)[28] to suggest that rising sea levels will inevitably exacerbate flooding conditions from storm events in the nearer-future than initially projected by the IPCC's fifth assessment report, leading to a linear increase in mean sea-level of 0.29 m by 2050 [27,28]. When considering a quadratic fit of these data, the curve suggests an elevated trend of 0.49m by 2050 [28]. Cities, counties, town governments, local institutions, and private contractors, provide myriad solutions, each of which must be evaluated in its own way, and the subsequent presentation of their flood data ultimately impact their efficacy as a warning. Also, provision of these serviceable flooding solutions often impacts the availability of other services citizens rely upon. Thus, this establishment of StormSense's flood monitoring blueprint is designed to aid other communities in mitigating adverse inundation impacts.

ACKNOWLEDGMENTS

The authors would like to thank Mike Ashe, Wade Gerze, Frank James, and the Newport News Public Works department and Oceaneering, Inc. for their efforts in installing, calibrating, and maintaining the water level sensors. The sensors were purchased from Valarm through funding assistance graciously provided by NIST and VDEM. We thank the reviewers whose conscientious comments improved this paper in many ways during the peer-review process. Portions of this publication and research effort are made possible through the help and support of NIST via federal award #70NANB16H277. This paper is Contribution No. 3734 of the Virginia Institute of Marine Science, College of William & Mary. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain commercial products are identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose.

LITERATURE CITED

[1] Rhee, S. 2016. Catalyzing the Internet of Things and Smart Cities: Global City Teams Challenge. SCOPE '16 Science of Smart City Operations and Platforms Engineering in partnership with Global City Teams Challenge (GCTC), p.1.

[2] Casulli, V. 2015. A conservative semi-implicit method for coupled surface-subsurface flows in regional scale, *International Journal for Numerical Methods in Fluids*, 79(4):199-214.

[3] Loftis, J.D., Molly Mitchell, Larry Atkinson, Ben Hamlington, Thomas R. Allen, David Forrest, Teresa Updyke, Navid Tahvildari, David Bekaert, and Mark Bushnell. 2018. Integrated Ocean, Earth and Atmospheric Observations in Hampton Roads, Virginia. Marine Technology Society Journal, 52(2), (*In Press*).

[4] Loftis, J.D., Forrest, D., Katragadda, S., Spencer, K., Organski, T., Nguyen, C., and Rhee, S. 2018. StormSense: A New Integrated Network of IoT Water Level Sensors in the Smart Cities of Hampton Roads, VA. Marine Technology Society Journal, 52(2), (*In Press*).

[5] Loftis, J.D., Wang, H.V. & Forrest, D.R. 2016. Street-Level Inundation Modeling of Hurricanes Matthew and Hermine and Emerging Flood Monitoring Methods in Hampton Roads. *William & Mary Publish*. URL

[6] Loftis, J.D., Wang, H.V. & Forrest, D.R. 2017. Catch the King Tide with StormSense on Nov. 5th: How You Can Help Crowd-Source Tidal Flood Event Calibrations for Hampton Roads' Newest Water Level Sensors. *William & Mary Publish*. URL

[7] Loftis, J.D., Wang, H., Forrest, D., Rhee, S., Nguyen, C. 2017. Emerging Flood Model Validation Frameworks for Street-Level Inundation Modeling with StormSense. SCOPE '17 Science of Smart City Operations and Platforms Engineering, 2(1), 13-18.

[8] Replicable Smart City Technologies (RSCT). 2016. Federal Funding Opportuninty Announcement: RSCT Cooperative Agreement Program. URL

[9] Sweet, W.V., Park, J., Marra, J.J., Zervas, C. & Gill, S. 2014. Sea level rise and nuisance flood frequency changes around the United States, in *NOAA Technical Report NOS COOPS 73*, 53 pp.

[10] Ezer, T. & Atkinson, L.P. 2014. Accelerated flooding along the US East Coast: on the impact of sea-level rise, tides, storms, the Gulf Stream, and the North Atlantic oscillations. *Earth's Future*. 2(8):362-382. URL

[11] Ezer, T. & Atkinson, L.P. 2017. On the predictability of high water level along the U.S. East Coast: can the Florida Current measurement be an indicator for flooding caused by remote forcing?, *Ocean Dynamics*, 67(6): 751-766.

[12] VanHoutven, G., Depro, B., Lapidus, D., Allpress, J. & Lord, B. 2016. Costs of Doing Nothing: Economic Consequences of Not Adapting to Sea Level Rise in the Hampton Roads Region. *Virginia Coastal Policy Center, College of William & Mary Law School Report*. URL

[13] Lawlor, J. 2012. City Line Apartments: flood prone and no solutions in sight. *Daily Press*, August 29, 2012. URL

[14] Alley, R.B. 2017. Letter regarding flooding on Newmarket Creek and City Line Apartments from Newport News Fire Chief. *Personal Correspondence*. URL

[15] Smith, H. 2016. After yet another City Line Apartments flood, FEMA steps in to help. *Daily Press*, October 12, 2016. URL

[16] FEMA. 2016. National Flood Insurance Program coverage isn't the same as homeowner insurance or FEMA assistance, Virginia Beach, Dec. 5, 2016. URL

[17] Bannan, B., Burbridge, J., Dunaway, M., Skidmore, D., Brooks, D., Crane, T., DiBiase, R., Dubrow, S., Hopingardner, P., Icenhour, T., Namuduri, K., Osborn, C., Phool, T.S., Purohit, H. & Thomas, G. 2017. Blueprint for Smart Public Safety in Connected Communities: *Initiative of the Global City Teams Challenge*. URL

[18] Wahl, T., Jain, S., Bender, J., Meyers, S.D. & Luther, M.E. 2015. Increasing risk of compound flooding from storm surge and rainfall for major US cities, *Nat. Clim. Change*, doi:10.1038/nclimate2736.

[19] Loftis, J.D., Wang, H.V., DeYoung, R.J. & Ball, W.B. 2016. Using Lidar Elevation Data to Develop a Topobathymetric Digital Elevation Model for Sub-Grid Inundation Modeling at Langley Research Center, *Journal of Coastal Research*, Special Issue 76:134-148.

[20] Loftis, J.D., Wang, H.V. & DeYoung, R.J. 2013. Storm Surge and Inundation Modeling in the Back River Watershed for NASA Langley Research Center, *NASA Technical Report,* NASA/TM-2013-218046.

[21] Loftis, J.D. 2014. Development of a Large-Scale Storm Surge and High-Resolution Sub-Grid Inundation Model for Coastal Flooding Applications: A Case Study during Hurricane Sandy, *Ph.D. Dissertation, College of William & Mary*. pp. 218.

[22] Wang, H., Loftis, J.D., Liu, Z., Forrest, D. & Zhang, J. 2014. Storm Surge and Sub-Grid Inundation Modeling in New York City during Hurricane Sandy. *Journal of Marine Sci. and Eng.*, 2(1), 226-246.

[23] Wang, H., Loftis, J.D., Forrest, D., Smith, W. & Stamey, B. 2015. Modeling Storm Surge and inundation in Washington, D.C., during Hurricane Isabel and the 1936 Potomac River Great Flood. *Journal of Marine Sci. and Eng*., 3(3), 607-629.

[24] Loftis, J.D., Wang, H.V., Hamilton, S.E. & Forrest, D.R. 2018. Combination of Lidar Elevations, Bathymetric Data, and Urban Infrastructure in a Sub-Grid Model for Predicting Inundation in New York City during Hurricane Sandy. *Computers, Environment, and Urban Systems* (*In Review*).

[25] Steinhilber, E.E., Boswell, M., Considine, C. & Mast, L. 2016. Hampton Roads Sea Level Rise Preparedness and Resilience Intergovernmental Pilot Project. *Phase 2 Report: Recommendations, Accomplishments and Lessons*. URL

[26] Loftis, J.D. 2017. "Catch the King" Tide Thank You and Review. CCRFR Thank You and Review Community Event at ODU, Dec. 13, 2017. Pres. 41. URL

[27] Mitchell, M., Hershner, C., Herman, J., Schatt, J., Mason, P., Eggington, E. & Stiles, W.S. 2013. Recurrent Flooding Study for Tidewater Virginia, *Report submitted to the Virginia General Assembly*, 135-150: URL

[28] Boon, J.D., Mitchell, M., Loftis, J.D. & Malmquist, D.M. 2018. Anthropocene Sea Level Change: A History of Recent Trends Observed in the U.S. East, Gulf and West Coast Regions. VIMS Special Report in Applied Marine Science and Ocean Engineering (SRAMSOE). No. 467. VIMS, College of William & Mary. URL

Nguyen, Cuong; Rhee, Sokwoo; Loftis, Jon 'Derek'.
"StormSense: A Blueprint for Coastal Flood Forecast Information & Automated Alert Messaging Systems."
Paper presented at Third International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE) in Partnership with Global City Teams Challenge (GCTC), Porto, Portugal. April 10, 2018 - April 10, 2018.

Cross-platform, Public Domain Simulation Tools for Performing Parametric IAQ and Energy Analysis

W. Stuart Dols
National Institute of Standards and Technology

Lindsay J. Underhill
Boston University School of Public Health, Boston, MA, U.S.A.

Engineering Laboratory, National Institute of Standards and Technology
100 Bureau Drive Gaithersburg, MD 20899

Presented at the
7th International Building Physics Conference
Syracuse, NY
September 23 – 26, 2018

U.S. Department of Commerce
Wilbur Ross, Secretary of Commerce

National Institute of Standards and Technology
Walter G. Copen, Under Secretary of Commerce for Standards and Technology and Director

NIST National Institute of Standards and Technology • U.S. Department of Commerce

Dols, William; Underhill, Lindsay.
"Cross-platform, Public Domain Simulation Tools for Performing Parametric IAQ and Energy Analysis."
Paper presented at 7th International Building Physics Conference, Syracuse, NY, United States. September 23, 2018 - September 26, 2018.

ABSTRACT

As building design is being driven towards lower energy use, the relationship between indoor air quality (IAQ) and energy becomes more important due in large part to reduced building envelope leakage, which can lead to higher indoor pollutant levels. Simulation tools that can analyze building design measures that aim to improve IAQ and energy use are necessary for evaluating potential trade-offs involving such measures. This paper will present the use of CONTAM and EnergyPlus, coupled using co-simulation, to perform parametric analysis of IAQ and energy impacts. Both of these tools are available in the public domain and provide cross-platform methods to evaluate both IAQ and energy use. Applications and workflow using these tools and available building models will be presented, including various energy and IAQ related measures that can be addressed with them. In particular, we present a framework for addressing energy measures (envelope tightening, insulation, and mechanical ventilation) and IAQ-related parameters (indoor/outdoor sources, ventilation rate, and filtration) in multi-family housing and effects on occupant exposure via a cohesive simulation environment that minimizes inter-domain coupling issues.

KEYWORDS

CONTAM, co-simulation, energy, EnergyPlus, indoor air quality, whole building simulation

INTRODUCTION

Building energy and indoor air quality (IAQ) are intertwined due to the interdependence of heat transfer, airflow, and contaminant transport. Often the same mechanical systems are utilized to maintain the thermal properties of air, e.g. temperature and relative humidity, and to dilute and/or remove pollutants that exist in the indoor environment, e.g., via outdoor air ventilation and filtration. As such, tools are needed to simulate these transport phenomena and associated systems to enable consideration of the interactions of these domains that are important to the health and comfort of building occupants. These tools will support the design and economic considerations of various stakeholders in the building community, including community planners, standards developers, designers and equipment manufacturers.

As highlighted by Teichman et al. (2015), activities related to design and construction of high-performance buildings (HPB) tend to focus heavily on energy-related concerns, and IAQ is often not addressed in a comprehensive and consistent manner. This is also borne out in the common use of building energy simulation, but not IAQ, in HPB design and analysis. However, recent activities by those evaluating HPBs from an IAQ perspective are bringing to bear building simulation methods that address both the energy and IAQ.

Building simulation is often employed to evaluate the impact of various building properties on building performance metrics (Azimi et al. 2016; Fabian et al. 2016). For example, improving building envelope airtightness can affect energy use, indoor contaminant concentrations, and occupant exposure. The ability to evaluate the myriad building types; heating, ventilating, air-conditioning (HVAC) systems; and climate zones, can provide information to those making decisions related to community-level energy use and contaminant exposure (Levy et al. 2016). To this end, two widely-used, public domain software tools, CONTAM and EnergyPlus, have been coupled to enable more complete evaluations of building performance (Dols et al. 2016). On their own, each tool is limited in its ability to account for transport processes upon which building IAQ, airflow and energy may be dependent.

EnergyPlus is a whole building energy simulation program with multizone heat balance as its underlying calculation method (Crawley et al. 2001). EnergyPlus determines zone thermal loads and the energy used by HVAC systems to meet those loads. It calculates zone air temperatures based on current system and plant capacity, including system airflow rates. Generally, infiltration and interzone airflows are user-specified, i.e., not pressure-dependent as in CONTAM, and are not required to be in balance with system airflow rates. Typically, infiltration is modelled based on correlations associated with rectangular, low-rise residential buildings or may be assumed to be constant, but better methods are available (Ng et al. 2018).

Dols, William; Underhill, Lindsay.
"Cross-platform, Public Domain Simulation Tools for Performing Parametric IAQ and Energy Analysis."
Paper presented at 7th International Building Physics Conference, Syracuse, NY, United States. September 23, 2018 - September 26, 2018.

CONTAM predicts airflows, contaminant concentrations, and airborne occupant exposures in multizone representations of whole buildings. In this paper, CONTAM will be used to assess IAQ while estimating infiltration airflows that impact building energy use. The CONTAM mass transport model treats a building as a system of interdependent zones or nodes (e.g., rooms, plenums and duct junctions) that store air and contaminant mass, and airflow paths (e.g., openings, cracks and duct segments) that transport air and contaminants between the nodes. Interzone airflows (including flows between the indoors and outdoors) are determined by calculating the node pressures that satisfy mass balance in each node based on driving forces and boundary conditions that include HVAC system airflows as well as wind and stack pressures exerted on the building envelope. CONTAM does not implement heat transfer calculations, so it requires indoor temperatures as inputs, which are often assumed to be ideally met thermostatic set-points.

The fact that CONTAM, when utilized on its own, requires the user to input zone temperature schedules, makes co-simulation with EnergyPlus an improved analysis approach. This is especially important for those who require analysis of both IAQ and energy related building performance.

METHODS

The National Institute of Standards and Technology (NIST) has been working with the Boston University School of Public Health to utilize co-simulation between CONTAM and EnergyPlus to evaluate the impact of energy retrofit programs in multi-family apartment buildings on energy savings and occupant exposure. Co-simulation is being used to evaluate multiple types of building energy retrofits, contaminant sources, and building ventilation systems.

Building Model Overview

The focus of the work to date has been on a four-story, mid-rise apartment building in Boston, Massachusetts. The Mid-Rise Apartment building model is based on the EnergyPlus representation selected from the set of U.S. Department of Energy (DOE) Commercial Reference Building models developed by the National Renewable Energy Laboratory (NREL) (Deru et al. 2011). NIST developed a corresponding CONTAM representation of this building (Ng et al. 2012) to be compatible with the co-simulation approach outlined in Dols et al. (2016). Both models were modified to include stair and elevator shafts that enable simulation of stack flows that can be particularly important to infiltration, energy use and contaminant transport in multi-story buildings.

The base building model, shown in Figure 1, consists of eight apartments on each floor separated by a central hallway with a stair and elevator shaft located at opposite ends of the hallway. Each apartment is served by a dedicated unitary HVAC system with a direct expansion cooling coil, a natural gas heating coil, and a constant volume supply fan. Each apartment is served by a dedicated exhaust system that is scheduled according to the ventilation system type: infiltration only, balanced outdoor air intake, or continuous exhaust ventilation.

Dols, William; Underhill, Lindsay.
"Cross-platform, Public Domain Simulation Tools for Performing Parametric IAQ and Energy Analysis."
Paper presented at 7th International Building Physics Conference, Syracuse, NY, United States. September 23, 2018 - September 26, 2018.

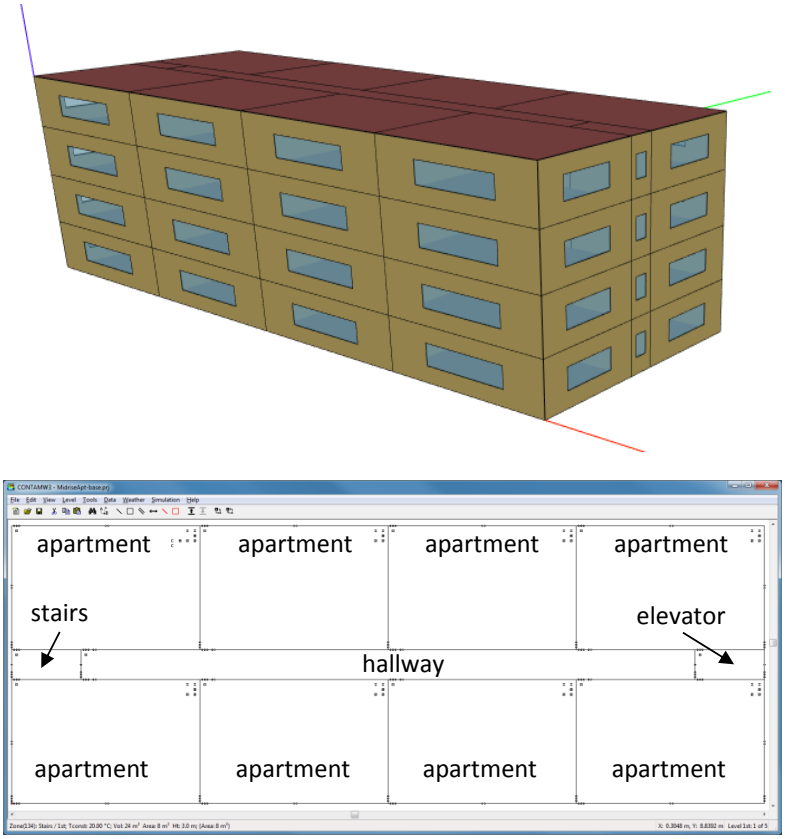


Figure 1. Mid-rise Apartment Building geometry (top) and floor plan in CONTAM (bottom)

Simulation Tools Development

The EnergyPlus/CONTAM co-simulation capabilities were previously developed as described in Dols et al. (2016). Coupling was implemented based on the Functional Mock-up Interface (FMI) for Co-Simulation specification according to which EnergyPlus was modified to enable the control of coupled simulations (Nouidui et al. 2013). However, the CONTAM co-simulation capability was originally implemented to execute only within the Windows operating system. To run a large set of parametric simulations, we required that the co-simulation capability be ported for execution on a high-performance, Linux cluster maintained by Boston University. EnergyPlus and the CONTAM simulation engine (ContamX) were already Linux compatible, so it was necessary to port the component that facilitates the FMI capability between EnergyPlus and CONTAM referred to as the ContamFMU dynamic link library (DLL). Modifications were made to enable the same source code to be used to build the Windows DLL (ContamFMU.dll) and the Linux equivalent referred to as a *shared object* (ContamFMU.so). Modifications were also required to address the methods used to spawn the ContamX process and enable socket communications to perform within the Linux, multi-core processing environment.

Simulation Setup

The simulation process and associated input files are illustrated in Figure 2. Base building models (template files) were developed for both EnergyPlus (IDF file) and CONTAM (PRJ file). Each of these templates was modified using a text editor to flag relevant values for replacement via a *Factorial Generator Tool* that reads both the flagged input file and a variable parameter file (*PRJ Parameters* and *IDF Parameters*) to create a full set of simulation input files. For the purposes of this demonstration case, Table 1 presents the set of parameters that were varied for a total of 810 simulations. However, these methods can be applied in an almost limitless number of combinations.

Dols, William; Underhill, Lindsay.
"Cross-platform, Public Domain Simulation Tools for Performing Parametric IAQ and Energy Analysis."
Paper presented at 7th International Building Physics Conference, Syracuse, NY, United States. September 23, 2018 - September 26, 2018.

The IDF files and PRJ files were generated by the *Factorial Generator Tool* prior to simulation. Scripts were then used to submit jobs to the process manager on the Linux cluster, after packaging files together as required for execution by EnergyPlus using co-simulation. The script then called EnergyPlus and CONTAM post-processing software (ReadVarsESO and simread3, respectively) to glean data from results files for further statistical evaluation.

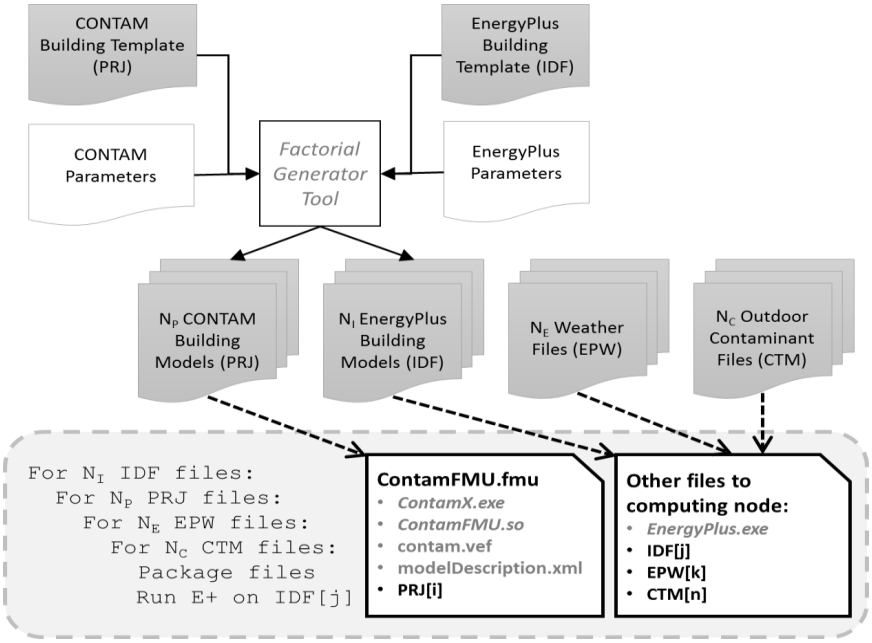


Figure 2. Schematic of parametric simulation process. N_P, N_I, N_E and N_C indicate number of respective file types: CONTAM building model (PRJ), EnergyPlus building model (IDF), weather (EPW), and outdoor contaminants (CTM).

Table 1. Set of Values for Parametric Simulations

Program	Parameter	Values
EnergyPlus (IDF file)	Ventilation Type	Infiltration only, Balanced supply, Exhaust
	Insulation (Walls/Roof)	R12/R13, R16/R30, R21/R35
CONTAM (PRJ file)	Envelope Leakage Rate (L/s·m^2 @75 Pa, exponent 0.65)	10.19, 5.42, 1.25
	Cooking Source	None, Low Cooking, High Cooking, Low Cooking w/ Local Exhaust, High Cooking w/ Local Exhaust
	Smoking Source	Non-Smoking, Smoking
	Filtration - Minimum Efficiency Reporting Value (MERV)	4, 8, 12

Dols, William; Underhill, Lindsay.
"Cross-platform, Public Domain Simulation Tools for Performing Parametric IAQ and Energy Analysis."
Paper presented at 7th International Building Physics Conference, Syracuse, NY, United States. September 23, 2018 - September 26, 2018.

RESULTS

Results presented here are based on simulations performed using Boston, MA weather and outdoor $PM_{2.5}$ data as described in Fabian et al. (2012), i.e., EnergyPlus weather (EPW) and CONTAM contaminant (CTM) files respectively. Detailed analysis of these results will be presented in future publications, but we present a subset of results to demonstrate the capabilities. The first case is a building with indoor particle sources of high-cooking activity and smoking, and outdoor particles, an indoor formaldehyde source, a MERV 4 filter in each air handler, and a relatively leaky building envelope. The second case is the same building with no indoor particle sources, MERV 12 filters, and a relatively tight building envelope. Each case was modelled with three types of ventilation systems: infiltration only, exhaust only, and balanced outdoor air. Figure 3 and Figure 4 present box-whisker data generated by CONTAM and show the average (line inside the boxes), standard deviation, and maximum and minimum air change rates and energy use (Figure 3) and concentrations averaged across all occupied zones (Figure 4).

Figure 3 shows whole-building air change rates and total annual energy use. In terms of energy use, all the buildings have the same insulation levels, so they only differ by envelope leakage and ventilation system type. As shown in Figure 3, the tighter buildings have reduced infiltration rates and lower total energy use for the respective ventilation systems.

Figure 3. Simulated whole building air change rates and total energy use of a Mid-rise Apartment Building with relatively leaky and tight building envelopes and three different ventilation systems: infiltration only (inf), exhaust only (exh), and balance outdoor air (oa).

Figure 4 shows indoor particle concentrations (grey boxes) and formaldehyde concentrations (yellow boxes). As expected, there are significant differences between indoor particle levels when source control and filtration are implemented. However, the indoor formaldehyde source leads to elevated concentrations in the tighter buildings especially when no mechanical ventilation is provided. Conversely, from the perspective of improving IAQ, increasing dilution by ventilation may reduce contaminant levels of indoor sources but could lead to increased levels of outdoor pollutants and increased energy use. This is demonstrated in the second case, which shows that exhaust only ventilation, when compared to infiltration only, has lower formaldehyde concentrations but slightly higher particle concentrations due to particles being drawn in through the building envelope, along with a higher total annual energy use. These examples highlight the need for an integrated approach to building design and analysis (ASHRAE 2017).

Dols, William; Underhill, Lindsay.
"Cross-platform, Public Domain Simulation Tools for Performing Parametric IAQ and Energy Analysis."
Paper presented at 7th International Building Physics Conference, Syracuse, NY, United States. September 23, 2018 - September 26, 2018.

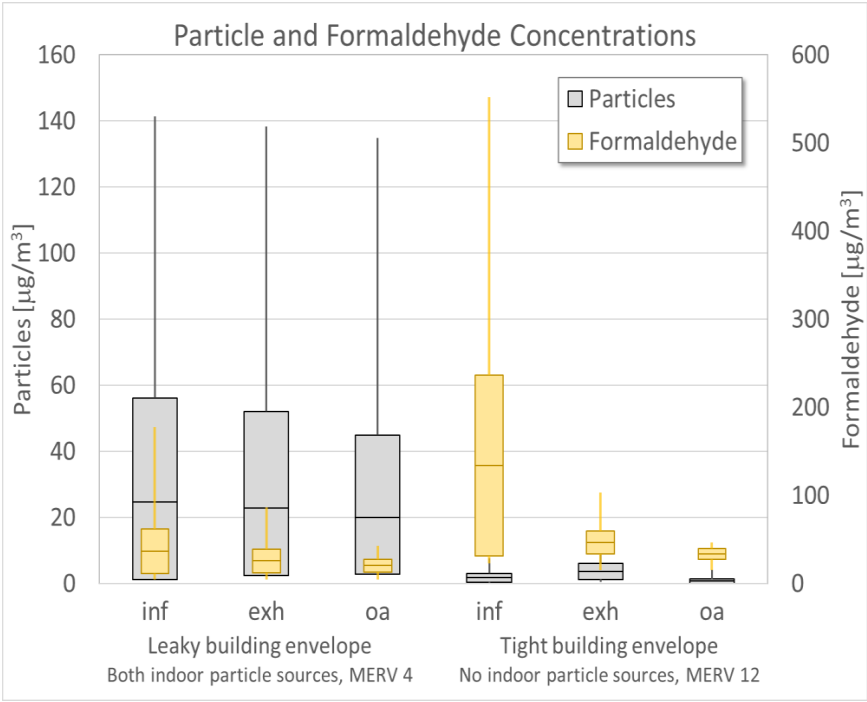


Figure 4. Simulated particle and formaldehyde concentrations averaged over all occupied zones of a Mid-rise Apartment Building for two cases of envelope air-tightness and source emissions scenarios with three different ventilation systems: infiltration only (inf), exhaust only (exh), and balance outdoor air (oa).

CONCLUSIONS

Co-simulation between whole-building IAQ, airflow, and energy simulation programs provides a comprehensive tool to evaluate the interactions between IAQ and energy when considering building energy retrofits. This paper highlighted the development and application of cross-platform, parametric simulation tools that provide the foundation to an integrated approach to building design and analysis to address energy and IAQ. The benefits of this integrated approach were demonstrated with an example from a case study carried out by Boston University and NIST. This example showed the interactions between building energy measures and IAQ parameters and their effects on whole-building energy use and occupant exposures, including reduced energy levels from envelope tightening and occupant exposure commensurate with source location and ventilation.

While these tools and parametric analysis methods are useful in their current state, there is also much work to be done to explore and improve them. For example, the coordination of CONTAM and EnergyPlus models is critical to success, and current methods rely on detailed knowledge of both simulation tools. Tools and associated workflows are available to minimize the redundancy and errors associated with coordinating the building representations, but modifications of existing building models can be quite cumbersome. Therefore, one of the goals of this work is to develop a set of coupled building models to be made publicly available. Output of the simulation tools can be voluminous and difficult to manage. However, they can also provide much greater insight into the interaction among the input parameters and building performance metrics than was presented herein. Work could also be done to provide outputs of desired metrics either directly or by enabling output to be easily manipulated by data processing utilities or scripts.

Dols, William; Underhill, Lindsay.
"Cross-platform, Public Domain Simulation Tools for Performing Parametric IAQ and Energy Analysis."
Paper presented at 7th International Building Physics Conference, Syracuse, NY, United States. September 23, 2018 - September 26, 2018.

ACKNOWLEDGEMENTS

The authors would like to thank Charles A. Jahnke of Boston University (BU) for his invaluable assistance in setting up the simulation platform on the BU shared computing cluster.

REFERENCES

ASHRAE. 2017. ASHRAE Position Document on Indoor Air Quality. edited by Refrigeration American Society of Heating, Air-conditioning Engineers. Atlanta, GA.

Azimi, Parham, Dan Zhao, and Brent Stephens. 2016. "Modeling the impact of residential HVAC filtration on indoor particles of outdoor origin (RP-1691)." *Science and Technology for the Built Environment*:1-32. doi: 10.1080/23744731.2016.1163239.

Crawley, Drury B., Linda K. Lawrie, Frederick C. Winkelmann, W. F. Buhl, Y. Joe Huang, Curtis O. Pedersen, Richard K. Strand, Richard J. Liesen, Daniel E. Fisher, Michael J. Witte, and Jason Glazer. 2001. "EnergyPlus: creating a new-generation building energy simulation program." *Energy and Buildings* 33 (4):319-331. doi: http://dx.doi.org/10.1016/S0378-7788(00)00114-6.

Deru, Michael, Kristin Field, Daniel Studer, Kyle Benne, Brent Griffith, Paul Torcellini, Bing Liu, Mark Halverson, Dave Winiarski, and Michael Rosenberg. 2011. US Department of Energy Commercial Reference Building Models of the National Building Stock. NREL/TP-5500-46861 National Renewable Energy Laboratory. Golden, Colorado.

Dols, W. Stuart, Steven J. Emmerich, and Brian J. Polidoro. 2016. "Coupling the Multizone Airflow and Contaminant Transport Software CONTAM with EnergyPlus using Co-simulation." *Building Simulation* 9:469-479. doi: 10.1007/s12273-016-0279-2.

Fabian, Maria Patricia, Sharon Kitman Lee, Lindsay Jean Underhill, Kimberly Vermeer, Gary Adamkiewicz, and Jonathan Ian Levy. 2016. "Modeling Environmental Tobacco Smoke (ETS) Infiltration in Low-Income Multifamily Housing before and after Building Energy Retrofits." *International journal of environmental research and public health* 13 (3):327.

Fabian, Patricia, Gary Adamkiewicz, and Jonathan I Levy. 2012. "Simulating indoor concentrations of NO2 and PM2. 5 in multifamily housing for use in health-based intervention modeling." *Indoor Air* 22 (1):12-23.

Levy, Jonathan I., May K. Woo, and Yann Tambouret. 2016. "Energy savings and emissions reductions associated with increased insulation for new homes in the United States." *Building and Environment* 96:72-79. doi: 10.1016/j.buildenv.2015.11.008.

Ng, Lisa C, Amy Musser, Andrew K. Persily, and Steven J. Emmerich. 2012. Airflow and Indoor Air Quality Models of DOE Reference Commercial Buildings. NIST Technical Note 1734. National Institute of Standards and Technology. Gaithersburg, MD.

Ng, Lisa C., Nelson Ojeda Quiles, W. Stuart Dols, and Steven J. Emmerich. 2018. "Weather correlations to calculate infiltration rates for U. S. commercial building energy models." *Building and Environment* 127 (Supplement C):47-57. doi: https://doi.org/10.1016/j.buildenv.2017.10.029.

Nouidui, Thierry, Michael Wetter, and Wangda Zuo. 2013. "Functional mock-up unit for co-simulation import in EnergyPlus." *Journal of Building Performance Simulation* 7 (3):192-202. doi: 10.1080/19401493.2013.808265.

Teichman, Kevin Y., Andrew K. Persily, and Steven J. Emmerich. 2015. "Indoor air quality in high-performing building case studies: Got data?" *Science and Technology for the Built Environment* 21 (1):91-98. doi: 10.1080/10789669.2014.976509.

1

Why Interoperability R&D Work Should be Driven by Agile Integration and Message Standards Concerns?[1]

Smart Manufacturing assumes agile integration of manufacturing services and applications [IVE 17, KUL 16]. This implies that manufacturing message standards, which are key to the services integration, need to be quickly created, used, and managed. That is a far cry from the best practices today. The interoperability R&D community should take on the challenge of enabling new message standards life-cycle management (LCM) capabilities needed for rapid integration and reconfiguration of manufacturing systems.

1.1. Vision: Rapid Integration and Reconfiguration of Manufacturing Systems

Manufacturing is on the cusp of being massively transformed: static and monolithic manufacturing systems are changed into dynamic networks of distributed manufacturing services. This transformation is even more radical now with the application of the Industrial Internet of Things (IIoT) technologies in manufacturing.

[1]Chapter written by Nenad IVEZIC and Boonserm KULVATUNYOU

Ivezic, Nenad; Kulvatunyou, Boonserm.
"Why Interoperability R&D Work Should be Driven by Agile Integration and Message Standards Concerns?."
Paper presented at 2018 Interoperability of Enterprise Systems and Applications (I-ESA), Berlin, Germany. March 22, 2018 - March 23, 2018.

In the new state, the required functionalities will be provided by services, which could be much more flexible and cost-effective to use. Manufacturing, however, requires sure-proof transformation approaches for the new service-based networks, including efficient integration, configuration, and re-configuration of services and IIoT solutions. Despite the complexity, visions of agile, re-configurable, service-oriented manufacturing systems continue to drive industry investments.

1.1.1. *Example: Smart Supply Chain Scenario*

Manufacturing Operations Management (MOM) or Enterprise Resource Planning (ERP) services of a manufacturing company will communicate with Third-Party Logistics (3PL) services to set up a temperature monitoring service in containers to carry the manufacturer's sensitive products. The messages necessary for setting up such service may include information about (1) the temperature range that will invoke a notification if there is violation, (2) what and how much data to send, (3) where to send the data, etc. Upon receiving a violation notification from a 3PL service (based on a communicated status of containers equipped with smart sensors) the MOM or ERP services will react to rapidly reconfigure the existing supply chain network. This will include placing additional order for the materials; rerouting or recalling the truck; dumping, and picking up new lot; etc.

1.2. Key Enabler: Message Standards for Services Integration

A key to the vision of smart manufacturing systems is that manufacturing services may be efficiently integrated, configured, and reconfigured to meet rapidly changing user requirements and market forces. Yet, the service integration and configurations depend on effective and meaningful service communications.

To provide for such services communications, correct message specifications must be developed for each specific use case (i.e., target business process). Message standards are used to develop such specifications in a precise and efficient manner.

A message standard refers to a set of shared rules and constraints, commonly called schema, that allow correct design and implementation of message exchanges and processing. Well-constructed message standards can enable efficient implementation of communicating services. This is a pre-requisite for efficient integration, configuration, and re-configuration of smart manufacturing systems.

Ivezic, Nenad; Kulvatunyou, Boonserm.
"Why Interoperability R&D Work Should be Driven by Agile Integration and Message Standards Concerns?."
Paper presented at 2018 Interoperability of Enterprise Systems and Applications (I-ESA), Berlin, Germany. March 22, 2018 - March 23, 2018.

1.3. A Major Problem: Simplistic Practices for Message Standards Development, Use, and Management

Today, a message standard (as a schema) may be very large with hundreds of thousands of data elements. Also, message standards are typically designed and maintained in an implementation-specific manner (e.g., with specific expression syntax and integration patterns), making it difficult to reuse the standard for a new implementation. Additionally, message standards management is very inefficient, often taking months to bring necessary updates to the user community.

These challenges in the current practice exist because message standards management (i.e., development, use, and maintenance) is a very complex affair, yet very poorly supported, leading to simplistic standards management practices. One consequence of these simplistic practices is that message standards are used to represent many use cases. Following such practices, a message schema is incrementally developed in a manner that ignores the intended usage. When the time comes for its use, the message schema is again used in a fashion that ignores any prior usage and it is used for each integration use case independently.

In such simplistic message schema development, the main principle is not to affect backward compatibility. This means that schemas are monotonically growing with all previously added components remaining within the new releases of message schemas. At the same time, the specific usage situation of the message standard additions, which drove the previous incremental change, is discarded. Message standards architects increment message standards with little concern to the previous usage situations. This results in limited re-use of previously added components and extensive additions of new components, giving rise to bloated message standards.

The bloated standards with poor usage documentations extend the problem to the usage phase. The users typically adapt the message standards independently for each specific and isolated use case. Such approach views a usage situation as a separate, independent episode of a message standard customization. Consequently, there is a very limited chance for meaningful reuse of standard components and virtually no chance of quickly achieving interoperable services using the message standards.

These complexities are even more evident for the new generation of dynamic and flexible networks of manufacturing services, which are designed for increased agility. There, the customization of message standards needs to happen very efficiently, as user requirements and market forces rapidly change, while new services are introduced in the ecosystem of providers and users of these services.

Ivezic, Nenad; Kulvatunyou, Boonserm.
"Why Interoperability R&D Work Should be Driven by Agile Integration and Message Standards Concerns?."
Paper presented at 2018 Interoperability of Enterprise Systems and Applications (I-ESA), Berlin, Germany. March 22, 2018 - March 23, 2018.

1.4. Root Cause: Lack of Support for Message Standards Life-Cycle Management (LCM)

If the message standards are to support efficient reconfiguration of networks of manufacturing services, an unprecedented support of message standards life-cycle management (LCM) is required. Such support needs to address concerns not dealt by the message standards development organizations (SDOs) to any significant extent: traceability, role-based management, collaboration, and continuity of LCM.

Traceability of LCM. The message-standards LCM needs to keep track of the customizations to the message standard to an unprecedented degree of precision. These customizations result in adaptations of message standards for a specific usage situation. That may introduce variability at any level of (i) granularity, (ii) aggregation, (iii) contextualization, and (iv) parallel customizations of components of message standards. Granularity- and aggregation-related traceability needs to track changes done on every level – from simple data types to complex message schemas. Contextualization-related traceability needs to track changes done for any usage situation – from general contexts to very specific contexts. A general context is defined by, say, type of manufacturing process and geographic region. A specific context is defined by, say, role, application, and event types for the message usages. Parallel customization-related traceability needs to track changes for numerous parallel usages of the standard in an ecosystem of service providers and users.

Role-based LCM. First, there is a need to support concurrent and independent customizations of the message standards by individual companies, which take on specific trading partner roles within various business processes. Second, there is a need for timely, synchronized management of message standard updates and releases by responsible standards development organizations.

Collaboration in LCM. First, as concurrent and independent customizations of a message standard take place in a specific company, multiple participants within the company will collaborate to adapt the message standard to their needs. Second, when management of a message standard release takes place at a standards development organization, the organization will collaborate with companies to manage customizations as well as new integration requirements into the standard.

Continuity of LCM. When the customizations of the message standards take place in companies, they will need to consider both prior customizations as well as prior and current releases of the message standard for its impact on existing and future operations of the company. In addition, when maintenance of a message standard release takes place at a standards development organization (SDO), it will

Ivezic, Nenad; Kulvatunyou, Boonserm.
"Why Interoperability R&D Work Should be Driven by Agile Integration and Message Standards Concerns?."
Paper presented at 2018 Interoperability of Enterprise Systems and Applications (I-ESA), Berlin, Germany. March 22, 2018 - March 23, 2018.

need to consider prior releases, customizations, and current release of the standard from the perspective of impacts on the existing and future users of the standard.

Today, there is a lack of tools that can provide the required traceable, role-based, collaborative, and continuous message standards LCM. This prevents rapid integration and configuration of communicating services and IIoT components.

1.5. Impact of the Lacking Message Standards LCM

If the message standards LCM is not traceable, role-based, collaborative, and supporting continuous processes, there will continue to be many undesirable consequences, preventing the efficient reconfigurability of manufacturing services.

With respect to *traceability* of LCM, while customizations of message standards have been provided at any level of granularity or aggregation, such customizations, in some cases, could not be captured with required precision. For example, differences in the intended usage of data types or elements at multiple places in a message component were not captured precisely, causing integration problems.

The lack of approaches to formally and commonly document contextualized use of a message standard results in limited reuse of message customizations, difficult-to-achieve interoperability, and bloated (i.e., very large and with redundant components) message standards. This makes message standards use inefficient. Cost of a message development is much higher then what may be possible if contextualized use were harnessed. Also, the systems integrations are brittle, with small requirements changes leading to unwanted behavior of the integrated system.

The lack of approaches to capture changes in a message standard, as they occur in parallel customizations, results in a failed opportunity to manage message standards in a proactive manner. Since message standards are large, and standards users are concerned only with a specific task of integration, there is a need to capture and act on independent, parallel changes in a proactive and constructive manner. Without such support, the standards will continue to be very hard and costly to use, especially in the open ecosystems of services with large parallel uses of standards.

With respect to *role-based* LCM, there is no distinction between a standards user and developer. A user has no support to manage multiple revisions of a single component type for multiple use cases, while a developer has no support to work on multiple releases that consider various standards customizations and requirements. These limitations result in inefficient standards development and use processes.

Ivezic, Nenad; Kulvatunyou, Boonserm.
"Why Interoperability R&D Work Should be Driven by Agile Integration and Message Standards Concerns?."
Paper presented at 2018 Interoperability of Enterprise Systems and Applications (I-ESA), Berlin, Germany. March 22, 2018 - March 23, 2018.

Collaborative capabilities, both within companies, and between companies and SDOs, are very limited and in the form of out-of-band, free-form messages. That, however, causes limited precision of, and low efficiency in, collaborative activities between various roles in collaborating organizations.

With respect to *continuity* of LCM, little exists in the way of support. Publishing a new release is largely a manual affair. No explicit relationships between elements, components, and types of the new and prior releases or customizations are kept. That causes limited understanding of the impact of changes. An existing customization of a message standard that needed to be moved to a new standard release needs to be manually analyzed and updated. This, in turn, is very tedious to accomplish on a realistic scale. Lack of such support prevents any automation when implementing the changes in message standards. Ultimately, updates to a message standard used in existing service configurations may be a very costly proposition.

1.6. Conclusion

Today, message standards development, use, and management take place over many weeks and months. That prevents efficient standards use for service-oriented, modularized applications integrations at a realistic scale, such as in enterprise-wide rationalizations of application functionalities and in an open marketplace of apps and services. In addition, the maintenance process of message standards typically continues indefinitely, presenting even greater challenges to the best practice processes. To enable reconfigurable services, we need to deliver new message standards life-cycle management capabilities. Development of technologies necessary for these capabilities is a challenge that the interoperability R&D community could take on and deliver impactful results to the industry.

1.6. Disclaimer

Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.

1.7. References

[IVE 17] Ivezic N., Kulvatunyou B., Brandl, D., Cho H., Lu, Y., Noller D., Davis J., Wuest T., Ameri, F. "Drilling down on Smart Manufacturing–enabling composable apps", *US Department of Commerce, National Institute of Standards and Technology*, https://doi.org/10.6028/NIST.AMS.100-8, 2017.

Ivezic, Nenad; Kulvatunyou, Boonserm.
"Why Interoperability R&D Work Should be Driven by Agile Integration and Message Standards Concerns?."
Paper presented at 2018 Interoperability of Enterprise Systems and Applications (I-ESA), Berlin, Germany. March 22, 2018 - March 23, 2018.

[KUL 16] Kulvatunyou B., Ivezic N., Morris, K.C., Frechette, S., "Drilling down on Smart Manufacturing–enabling composable apps", *Manufacturing letters*, vol. 10, 2016, p. 14-17.

Ivezic, Nenad; Kulvatunyou, Boonserm.
"Why Interoperability R&D Work Should be Driven by Agile Integration and Message Standards Concerns?."
Paper presented at 2018 Interoperability of Enterprise Systems and Applications (I-ESA), Berlin, Germany. March 22, 2018 - March 23, 2018.

Proceedings of the 2018 International Mechanical Engineering Congress and Exposition
IMECE2018
November 9-15, 2018, Pittsburgh, PA USA

IMECE2018-87686

VIRTUAL EXPERIMENTAL INVESTIGATION FOR INDUSTRIAL ROBOTICS IN GAZEBO ENVIRONMENT

Murat Aksu
National Institute of Standards and Technology
Gaithersburg, Maryland USA

John L. Michaloski
National Institute of Standards and Technology
Gaithersburg, Maryland USA

Frederick M. Proctor
National Institute of Standards and Technology
Gaithersburg, Maryland USA

ABSTRACT

Measuring the agility performance of the industrial robots as they are performing in unstructured and dynamic environments is a thought-provoking research topic. This paper investigates the development of industrial robotic simulation algorithms for the effective application of robots in those changing environments. The distributed framework for this investigation is the Robot Operating System (ROS) which is extensively used in robotic applications. ROS-Industrial (ROS I), which extends the capabilities of ROS to manufacturing, allows us to interoperate between industrial robots, sensors, communication buses and other kinds of automation tools. Gazebo is used as the open-source 3D simulator to design a virtual industrial robotic system, which is a prevailing tool as a node in the ROS environment. An effort is underway to replicate the in-house experimental robotic kitting lab with a graphical physics simulation that can be shared worldwide. This graphical physics simulation is not tied to a specific robotic control system. An experimental approach will be presented detailing the issues related to a physics based simulation of kitting with multiple collaborative robots, multiple tools, parts, tool changers, safety system, and sensors. In this realm, the ability for the simulation environment to encompass the current system as well as additional more complex sensors and actuators will be discussed. To make this simulation environment more realistic, Gaussian noise will be introduced to the data generated by virtual sensors. We expect that this experimental approach will be a seamless way for users to verify and validate their control systems even if they do not have a physical robot at their facilities.

INTRODUCTION

Industrial robots have traditionally been applied to automate tasks that are dirty, dull, or dangerous, and have been a key driver of continued productivity growth in high-volume applications such as automotive and electronics manufacturing. Worldwide, there were about 1.8 million installed industrial robots at the end of 2016, with a 10 % annual growth rate since 2010 [1]. In these high-volume applications, long up-front programming times are acceptable. However, as manufacturing requires more flexibility to support quickly changing product requirements in a high-mix, low-volume environment, robots need to become more agile. Agility in this context refers to the ability of a robot to be rapidly re-tasked for new activities without being taken offline for programming, the ability of a robot to recover from errors or uncertainty in the environment, and the ability to move applications between robots from different vendors without the need for program translation.

We have focused on how to represent the knowledge needed to achieve robot agility, the system architecture and component integration, planning, sensing and control, and how to measure the agility performance of robotic systems [2]. In support of this work, we have relied on simulation as a tool for the development of robotics systems to find ways to make better-informed decisions. In our laboratory, simulation allows tests to be run without contending for scarce time on physical robots, and to conduct tests safely without risk to damaging robots and tooling or injuring people. Simulation has also been used as the basis for competitions on robot agility, enabling competitors to practice at their facilities and compete in a controlled and instrumented environment. These simulations have used the Gazebo physics-based simulation package [3],

Aksu, Murat; Michaloski, John; Proctor, Frederick.
"VIRTUAL EXPERIMENTAL INVESTIGATION FOR INDUSTRIAL ROBOTICS IN GAZEBO ENVIRONMENT."
Paper presented at 2018 International Mechanical Engineering Congress and Exposition (IMECE2018), Pittsburgh, PA, United States.
November 9, 2018 - November 15, 2018.

supplemented with the Robot Operating System (ROS) open-source framework for robot control [4].

There are several reported works aimed at simulation and robot control which are based on Gazebo and ROS. Aguero *et al.* [5] presented a Gazebo simulation platform for a cloud-hosted humanoid robot simulation with a wide range of sensors, controllers, and actuators to address the challenge of real-time task-oriented rescue robot competition for the Defense Advanced Research Projects Agency (DARPA) Virtual Robotics Challenge (VRC), and showed that simplified dynamics while maintaining sufficient accuracy is feasible. Swanson *et al.* [6] studied a Hardware-in-the-Loop (HIL) driving simulator that served as a training platform for driving performance, and presented a driver-in-the-loop simulation environment in which the driver and vehicle hardware components interacted with each other using ROS and Gazebo. The authors emphasized the effect of computer processor choices on decreasing the latency in the simulator and increasing the system fidelity. Fernandes *et al.* [7] analyzed the simulation of autonomous control of a robotic car using ROS and Gazebo, a research challenge of the Brazilian National Institute of Science and Technology on Embedded Critical Systems (INCT-SEC); however, very little information was provided on the driving simulation environment itself. Qian *et al.* [8] investigated the simulation of a robotic arm for manipulating objects by building a model of a pick-and-place robot with seven degrees of freedom, and demonstrated methods to implement robot control in a short period of time using ROS and Gazebo.

Unlike the research reported here, most of the simulation studies in the literature deal with non-industrial robotics applications. Some of the novel contributions presented in our research are a standard for sending commands and receiving status between an industrial robot and controller, a method of providing noisy sensor information to the object recognition system in our laboratory, and an overall effort to replicate the entire NIST agility laboratory with physics-based simulation to make it available to external collaborators allowing them to test their algorithms worldwide. The following sections will provide some background in the research, and describe how physics-based simulation of system components has helped in the research efforts.

Knowledge Representation

To help automate the planning of robot activities, a model of the robot's attributes, capabilities, and environment is needed. IEEE 1872, the Core Ontology for Robotics and Automation (CORA), is a standard for representing this information [9]. CORA provides definitions for general concepts for robotics, to support automated reasoning about robot activities, and as the basis for exchanging information about robotics.

Supplementing CORA, the authors have developed a messaging language for sending commands to robots and receiving real-time status from them. This messaging language, the Canonical Robot Command Language (CRCL) [10], is an eXtensible Markup Language (XML) Schema Definition (XSD) for information used to integrate and task robots independent of their internal programming language. Command message content includes:

- setting units, speeds, accelerations, and tolerances,
- setting robot parameters,
- performing Cartesian motions,
- performing joint-level motions,
- operating an end effector,
- configuring status reports,
- getting a status report immediately,
- displaying messages, and
- pausing or stopping motion.

Planning

A hallmark of agile robot systems is their ability to automatically plan and replan their activities in a dynamic and changing environment. To test this ability, the authors have used the Planning Domain Definition Language (PDDL) [11] as a source for specifying planning test inputs and results. PDDL is a language for encoding information needed for general planning problems, such as vehicle routing [11] and robot assembly [12]. Planners that use PDDL refer to models of objects, predicates about objects that can be true or false, the initial state and goal state of the world, and actions that are available to move objects and transform the initial state of the world to its desired goal state. These are placed into a domain file (for predicates and actions), and a problem file (for objects and world states). Given these files, a PDDL planner generates a series of actions to solve the planning problem. PDDL thus provides a standard way to measure the performance of planning algorithms in terms of time taken, memory used, quality of plan, or other metrics.

Figure 1 shows an excerpt from a PDDL domain file for a kitting application. The `:types` are tags for the resources referenced by the problem specification and the resulting plan file. The `:predicates` identify tags for resources whose state will be queried when evaluating the predicates. Predicate definitions are not shown, but are essentially lists of conditions to be evaluated by the application that executes the plan.

```
(define (domain kitting-domain)
  (:types
  EndEffector
  EndEffectorChangingStation
  EndEffectorHolder
  Kit
  KitTray
  LargeBoxWithEmptyKitTrays
  LargeBoxWithKits
  Part
  PartsTray
  Robot
  StockKeepingUnit
  WorkTable)
(:predicates
  ; part is held by endeffector
  (part-has-physicalLocation-refObject-endEffector
```


Aksu, Murat; Michaloski, John; Proctor, Frederick.
"VIRTUAL EXPERIMENTAL INVESTIGATION FOR INDUSTRIAL ROBOTICS IN GAZEBO ENVIRONMENT."
Paper presented at 2018 International Mechanical Engineering Congress and Exposition (IMECE2018), Pittsburgh, PA, United States.
November 9, 2018 - November 15, 2018.

```
    ?part - Part ?endeffector - EndEffector)
  ; parts tray contains part
  (partsVessel-has-part ?partstray - PartsTray
    ?part - Part) ... )
```

Figure 1. Sample PDDL domain for a robotic kitting application.

Figure 2 shows an excerpt from a PDDL problem file that is to be solved by a planner, to take the kitting application from a stating initial state `:init` to ending goal state `:goal`. The planner's objective is to determine a series of actions that take the system from the initial state to the final state as efficiently as possible.

```
(define (problem kitting-problem)
  (:domain kitting-domain)
  (:objects
    robot_1 - Robot ... )
  (:init
    (endEffectorHolder-has-endEffector
tray_gripper_holder part_gripper)
    (partsVessel-has-part  part_small_gear_tray
small_gear_1) ... )
  (:goal (and
    (= (final-quantity-of-parts-in-kit kit_s2b2) 4)
    (= (quantity-of-parts-in-kit
      sku_part_small_gear kit_s2b2)
       (capacity-of-parts-in-kit
      sku_part_small_gear kit_s2b2)) ) ) )
```

Figure 2. Sample PDDL domain for a robotic kitting application.

Figure 3 is a sample plan showing actions for locating, picking, and placing a part in one step of an overall larger plan for kitting.

```
(look-for-part robot_1 large_gear_1
  sku_part_large_gear kit_s2b2 part_gripper)
(set-grasp robot_1 large_gear_1
  sku_part_large_gear part_gripper)
(take-part robot_1 large_gear_1 sku_part_large_gear
  part_large_gear_tray part_gripper kit_s2b2)
(look-for-slot robot_1 large_gear_1
  sku_part_large_gear kit_s2b2 part_gripper)
(place-part robot_1 large_gear_1 sku_part_large_gear
  kit_s2b2 part_gripper work_table_1
  part_medium_gear_tray)
```

Figure 3. Sample PDDL plan.

As noted by Mösenlechner and Beetz [13], the logical PDDL model of actions that trigger known changes in state is not well suited to autonomous robot systems, where the outcome of actions may not be predictable. Another problem is that the high-level nature of PDDL states for the initial conditions, goal conditions, preconditions, and postconditions do not incorporate finer-grained detail whose slight variation could lead to different choices of actions. As described in the next section, the first shortcoming of PDDL has been overcome by incorporating continuous replanning when actions do not result in the predicted outcome. The second shortcoming has been addressed to by choosing actions with parameters at a level of resolution low enough be adjusted through real-time sensor feedback from vision.

System Architecture

Work supporting CORA and CRCL has taken place in the Agility Performance of Robotics System (APRS) laboratory at the National Institute of Standards and Technology (NIST) [2]. The lab contains two industrial robots, a Fanuc LR-Mate 200iD and a Motoman SIA20F. The robots share tooling for open-and-close gripping and vacuum gripping. The primary application is kitting, where parts are moved from their initial location in storage trays to a final target arrangement in kit trays. Overhead cameras in the work volume are used to determine the location of parts, storage trays, and kit trays. This laboratory is shown in Figure 4.

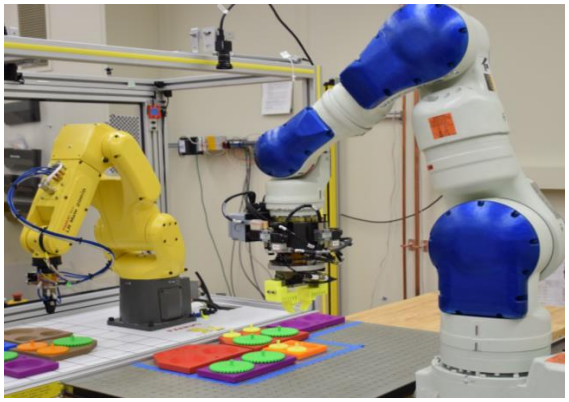


Figure 4. APRS Laboratory Workcell.

Figure 5 shows the APRS system architecture. The purpose of the system is to put together kits of parts based on a request composed by the operator, shown at the top of the figure. The resulting PDDL goal is a set of kits and their contents of parts. This goal is sent to the PDDL planner, which consults the definitions of actions, preconditions, and postconditions in the kitting problem domain and determines a feasible sequence of actions that achieve the kitting goal. These actions are sent to the PDDL executor, which fills in actual values for part and kit locations based on the current state in the world model database. This database is continually updated with the locations of parts as measured by the object recognition system. After the actions are instantiated, the resulting CRCL program is sent to the CRCL client for execution. This client steps through the program, sending messages to the robots and grippers and monitoring execution status until the program has completed. Any failures are reported by the client, which triggers replanning if possible until the kitting request is fulfilled, or stopped due to unrecoverable problems.

Aksu, Murat; Michaloski, John; Proctor, Frederick.
"VIRTUAL EXPERIMENTAL INVESTIGATION FOR INDUSTRIAL ROBOTICS IN GAZEBO ENVIRONMENT."
Paper presented at 2018 International Mechanical Engineering Congress and Exposition (IMECE2018), Pittsburgh, PA, United States.
November 9, 2018 - November 15, 2018.

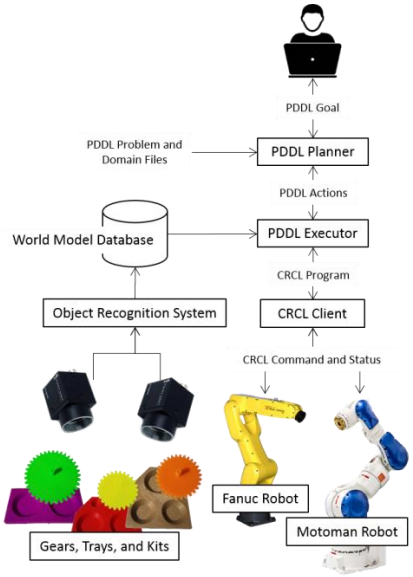


Figure 5. System Architecture.

Measuring Agility

To advance the state of robot agility, a series of competitions was organized that measure the effectiveness of planning systems to rapidly re-task robots without the need for human intervention. Such tasking includes the ability of robots to recover from errors such as dropped parts, the ability of perception systems to identify problems, and the ability of manipulation systems to reposition objects for better error recovery. The Agile Robotics for Industrial Automation Competition (ARIAC) is sponsored by NIST in collaboration with the Open Source Robotics Foundation (OSRF), developers of the Robot Operating System (ROS) [4] and the Gazebo physics-based simulation environment.

Competitions are organized around a kitting application, where robots are given the task to move objects from a set of trays to a goal kit. Virtual sensors for determining object locations include cameras, beam break detectors, laser range scanners, and laser line curtains. Teams are given the flexibility to choose which sensors to use. Costs are associated with the sensors selected and factored into the scoring metrics.

THIS RESEARCH

The goal of this research is to use a physics-based simulation to stand in for the APRS laboratory environment, and use the uncertainty in part location and activity completion to test the ability of the planning system to recover from failures. The intent of the simulation is to provide the following enhancements:

- the ability to test strategies for sensor-based recovery from errors in a repeatable environment;
- enabling hybrid real-virtual operation, where one real robot and camera and one simulated robot and camera can be used simultaneously;
- and to provide additional implementations of CRCL-conforming robots to validate this specification.

Gazebo was selected as the simulation environment [3]. Gazebo provides realistic visual rendering of physical scenes, linked to one of a set of configurable physics engines that update the state of objects in the simulated world according to physics principles such as friction, inertia, and gravity.

Figure 6 shows a Gazebo visualization of the physics-based simulation of a kitting activity used in the ARIAC competition.

Figure 6. Gazebo Simulation of Robot Kitting in the ARIAC Competition.

Simulating the APRS Environment

The NIST agility kitting robot control laboratory was ported to a physics-based simulation environment. Simulation can be kinematic or physics-based. Kinematic simulation uses visualization of the sequence of operations to verify correctness. Physics-based simulation models the physical elements' interactions and collisions and the effects of physical properties such as gravity, friction, and inertia. The intent of physics-based simulation is to study control and sensing, reveal inaccuracies, and verify correctness. For example, placing of a "gear" into a slot holder in a visualization could overlay two images at the bottom of the slot (the gear and holder) without repercussions. However, in the case of physics-based simulation, the gear would "bounce" out of the slot as it is physically impossible for a solid object to atomically combine with another solid object.

Figure 7 shows the Gazebo physics-based simulation of the agility lab. The simulation modeling includes the two robots, the agility lab physical space, which consists of the tables, the walls, and finally the gear, kitting, and tray objects. The two overhead vision cameras and enclosing safety system are not visualized in the simulation; however, the camera images are simulated by a Gazebo plugin that can be used to test the actual robot planning and control system. Robotiq two-finger gripper models available on the Internet were used, saving the effort of

Aksu, Murat; Michaloski, John; Proctor, Frederick.
"VIRTUAL EXPERIMENTAL INVESTIGATION FOR INDUSTRIAL ROBOTICS IN GAZEBO ENVIRONMENT."
Paper presented at 2018 International Mechanical Engineering Congress and Exposition (IMECE2018), Pittsburgh, PA, United States.
November 9, 2018 - November 15, 2018.

modeling the custom 3D-printed fingers used in the lab. Grippers and the robot base location are handled in a kinematic ring (discussed later), which makes substitution of different grippers and relocation of the robots in the agility lab done with a different transform.

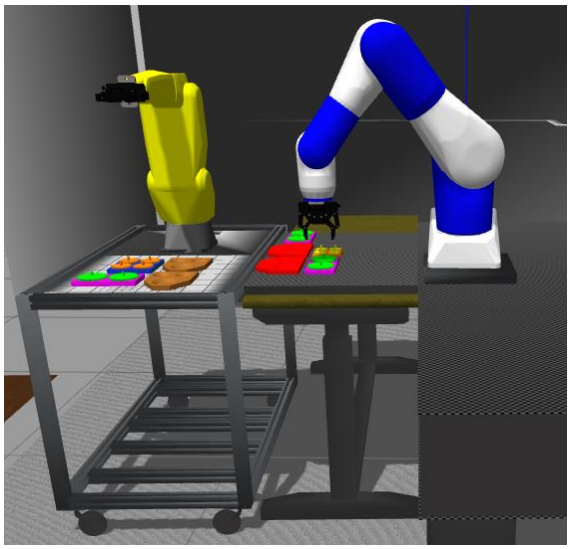


Figure 7. Agility Lab Physics-Based Simulation

Modeled objects in the simulation environment.

The Gazebo physics-based simulation relies on either Gazebo Simulation Description Format (SDF) or ROS Universal Robotic Description Format (URDF) to model robots and other world elements. Both SDF and URDF are XML file formats that describe objects and environments for robot simulators, visualization, and control [14, 15]. Both URDF and SDF include mechanisms to describe links, joints, kinematic chain relationships, the limits and capabilities of the joints, obstacle volume of a link, and a visual representation of each link. SDF includes mechanisms to describe physical elements such as mass, inertial frame, gravity interaction, among a multitude of physics descriptors. Often, ROS URDF was converted into Gazebo SDF format. Thus, in the physics-based agility simulation, the robot and gripper, the kitting objects, and the agility lab were defined with either SDF or URDF.

For example, the agility simulation includes gears, holders, and kits that were originally modeled for 3D printing, but the Computer Aided Design models were translated into STL format (i.e., stereolithography), and modified to satisfy the Gazebo world (for example, adding mass and inertial frames while adjusting the origin coordinate frame).

Robot Control

The agility robot control application domain is multi-axis, coordinated motion control. In addition, process control is necessary to handle input/output, and auxiliary equipment. Representative robot controller applications include manipulation, assembly, and collaboration.

The robot controller software modules include: (1) Joint control, which performs servo control of axis motion by transforming incoming motion goal points into set-points for the corresponding actuators, (2) Cartesian motion planning, which coordinates the motions of an individual joint by transforming an incoming motion segment specification into a sequence of equi-time-spaced setpoints for the coordinated axes, and (3) Task Coordinator modules, which sequence operations and coordinate the various motion, sensing, and event-driven control processes. Overall, the robot control architecture was based on open-source components. Its source code is available for public use.

The inverse and forward kinematics use OpenRAVE IKFast software to solve the forward and inverse kinematics [16]. In general, IKFast can analyze the robot kinematics, solve the kinematics equations, and write the solution to a C++ file.

Because the NIST agility robot motion control relies on Cartesian based straight-line motion and assumes a collision-free industrial environment, trajectory planning is done with "Gotraj" [17]. Gotraj computes a smooth trajectory based on either time or dynamical properties (velocity, acceleration, and jerk). Gotraj assumes a final velocity of zero. Users can append poses onto the Gotraj motion queue that will result in additional trajectories. Gotraj supports a "stop" motion directive that will generate a trajectory which will stop as soon as possible given the motion dynamical properties (velocity, acceleration, and jerk).

Another common robotic concept that was required to handle different robots, grippers, and robot locations within the simulation world was the kinematic ring. Thus, any CRCL commanded robot Cartesian motion in world coordinates is then expressed in terms of kinematic ring made up of a series of homogeneous matrix transforms from the base robot frame, including the robot transformation, and robot gripper transform. With this mechanism, it was easy to shift between world and robot coordinate frames. This is important as although the robot and kitting objects are modeled precisely, the location of the robot base and the gripper tooling can vary, and kinematic rings offer a convenient scheme for resolving the variances.

Integration with CRCL. At NIST, kitting commands to the robot(s) are expressed in CRCL, which is a messaging language for controlling a robot that is executed by a low-level device robot controller. CRCL is an abstraction of the robot control capabilities that is fully defined in an XML schema. CRCL contains the ability to command robot joint, Cartesian position and gripper control. In addition, CRCL supports status streaming or service requested communication patterns. CRCL uses XML messages for communication, which typically uses a stream based Transmission Control Protocol (TCP) socket to communicate the XML.

Aksu, Murat; Michaloski, John; Proctor, Frederick.
"VIRTUAL EXPERIMENTAL INVESTIGATION FOR INDUSTRIAL ROBOTICS IN GAZEBO ENVIRONMENT."
Paper presented at 2018 International Mechanical Engineering Congress and Exposition (IMECE2018), Pittsburgh, PA, United States.
November 9, 2018 - November 15, 2018.

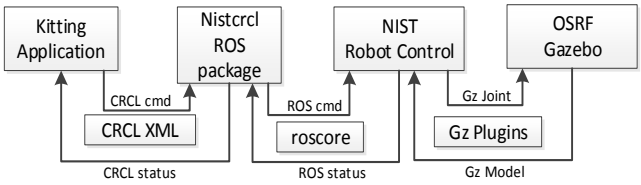


Figure 8. ROS Nistcrcl Package Architecture.

Figure 8 shows the ROS Nistcrcl package architecture used to handle CRCL communication in the physics-based robot agility simulation. The ROS Nistcrcl package is a ROS node that was developed to listen and broadcast CRCL XML command and status messages, while translating CRCL from/to ROS representation and communication ideology. The Nistcrcl ROS package uses several open-source software technologies to adapt CRCL into ROS. The CodeSynthesis XSD tool generates a C++ object model from XSD that is used to parse and serialize CRCL XML. CodeSynthesis relies on the Xerces XML parser. The Boost C++ library Asynchronous IO (Asio) was used to handle socket command and status communication with the CRCL client. The ROS message infrastructure (based on the Google protobuf open source communication scheme) was used to build "custom" ROS topics that encapsulated CRCL functionality. Two ROS custom messages were developed - one for CRCL command messages and one for CRCL status messages.

Gripping objects. In general, an end effector is the device at the end of a robotic arm, meant to interact with the environment. Kitting is concerned with grasping objects and relocating the objects. This can be done with a vacuum gripper, or a gripper with two or more fingers. We are interested in the case of using grippers to achieve object manipulation (i.e., grasping and releasing).

One gripper used for simulation was the Robotiq's 2-Finger Adaptive Robot Gripper. An open source ROS URDF description existed to describe the kinematics and visualization, which simplified implementation.

Another gripper used in the simulation was the NIST in-house 3D printed parallel jaw gripper. Grasping objects highlights the difficulty of physically modeling gripper and grasped object interaction, because explicit enumeration of friction, collision, and other dynamical behavior elements play an important role in grasp control. By comparison, in the real world, grasping generally ignores the interaction of the underlying physics.

Calibration

The APRS has four coordinate systems: the Fanuc and Motoman robots, and the two cameras located above their work volumes.

Calibration of camera coordinates to robot coordinates is done following a four-point registration procedure:

1. A marker object (e.g., a gear) is placed at a location in the work volume of one of the robots.
2. The object detection system's coordinates for the center of the marker object are recorded.
3. An operator guides the robot to the center of the marker object, and the robot controller's coordinates of this point are recorded.
4. This procedure is repeated for a total of four locations, placed near the corners of the work volume.
5. Offset transformations from object detection system coordinates to robot coordinates are computed at each of the four locations.

These offset transformations are determined only occasionally, when the robots or cameras are repositioned. Table 1 shows the results of four-point registration for the Motoman robot.

Table 1. Calibration Offsets from Object Recognition System to Motoman Robot.

X (mm)	Y (mm)	X offset (mm)	Y offset (mm)
535.9	157.7	0.0	5.0
723.1	-400.3	-9.0	-7.0
523.8	-106.7	-4.0	0.0
747.3	155.4	5.0	0.0

During operation, object locations from the object detection system are interpolated between these four registration transforms, resulting in a location of the object in robot coordinates. This interpolation is done by weighting each of the contributions of the registration points by their inverse distance to the point whose offsets are to be interpolated, normalized by the sum of the inverse distances. The problematic infinities for the values of the inverse distances at the registration points themselves are avoided by algebraically changing the formula to that in Eqn. (1):

$$\delta_p = \frac{\sum_{i=1}^{n} \delta_i \prod_{j \neq i}^{n} d_j}{\sum_{i=1}^{n} \prod_{j \neq i}^{n} d_j} \qquad (1)$$

where δ_p is the interpolated offset, the set of δ_i are the offsets at the n registration points, and set of d_j are the distances from the point to be interpolated to each of the registration points. This equation works for offset translation vectors as well as offset orientations, when orientations are represented as rotation vectors.

The four-point registration procedure is also used to determine the transformation between the Fanuc and Motoman robot coordinate systems. Using this procedure, a best-fit transform is computed following Horn's solution to the absolute orientation problem [18].

Inverse error compensation. It is desirable for the simulated behavior to match the real-world behavior, so that the four-point registration procedure need not be repeated in simulation and consequently require the control system to

Aksu, Murat; Michaloski, John; Proctor, Frederick.
"VIRTUAL EXPERIMENTAL INVESTIGATION FOR INDUSTRIAL ROBOTICS IN GAZEBO ENVIRONMENT."
Paper presented at 2018 International Mechanical Engineering Congress and Exposition (IMECE2018), Pittsburgh, PA, United States.
November 9, 2018 - November 15, 2018.

switch between real and simulated operation. Therefore, the simulated object detection system needs to adjust its output with the inverse of the interpolated transform. Figure 9 shows a map of the magnitude of the compensating error throughout a region bounded roughly by the registration points.

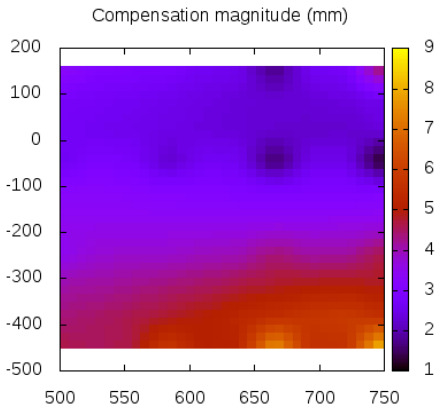


Figure 9. Inverse Error Compensation

The inverses of these positional compensation errors will be applied to the reporting of simulated object locations, as detailed in the following sections.

Object Recognition and Reporting

There are two methods for recognizing objects: ground truth with noise, and camera image generation with noise. For the first method, the computer vision system is not used, and its output (object locations) is simulated directly. For the second method, synthetic camera images are produced from the simulation, and fed to the computer vision system instead of images from real cameras. The goal of the simulation is to be able to provide realistic object data streams to the CRCL executor, without reconfiguring the executor based on whether real or simulated sensor data is used. Figure 10 shows view of the objects in the real object detection system.

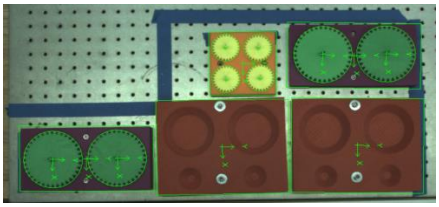


Figure 10. Objects Appearing in the Object Detection System.

The output characteristics of the object recognition system were first determined based on a test of its measurements of a single object. Figure 11 shows a plot of the X-Y positions of 1487 measurements of the position of a single object by the object recognition system. The figure shows clustering of points and does not suggest any well-known underlying distribution. For the purposes of simulating randomness consistent with the variation shown, a normal distribution was fit to the data, with a mean and standard deviation of $\mu = 416.65$ mm and $\sigma = 0.28$ mm for X, and $\mu = 342.50$ mm and $\sigma = 0.33$ mm for Y. The following sections describe how the simulation was developed so that it produces sensor data consistent with that produced by the actual sensing system.

Ground truth with noise. Gazebo provides the true locations of all objects in the world through a ROS topic updated at the simulation frequency, measured to be about 2 milliseconds. These ideal locations of the gears, kits, and trays are fed into an application that adjusts the poses of each object with noise representative of that measured by the real object detection system.

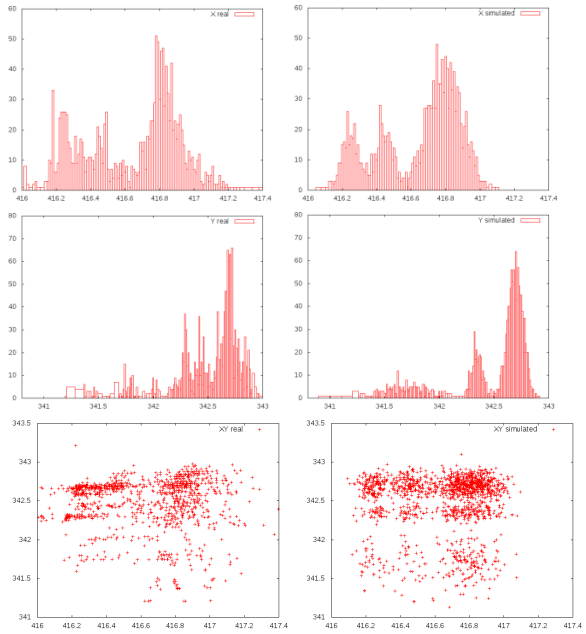


Figure 11. X-Y Location Variance from Object Recognition System, Real and Simulated. Left Images Are Real, Right Images Simulated.

The noise-adding application was customized so that it replicates the clustering exhibited by the real object detection system. For each of the X and Y distributions shown in Figure 11, three normal distributions were composed with means, standard deviations, and relative contributions that approximated the clusters. This empirical customization, while not perfect, replicates the behavior of the system to a degree that is visible to the CRCL executor and has the same qualitative influence.

Camera image generation with noise. With this method, the existing object recognition is used, and synthetic camera images are provided to it as if they originated from

Aksu, Murat; Michaloski, John; Proctor, Frederick.
"VIRTUAL EXPERIMENTAL INVESTIGATION FOR INDUSTRIAL ROBOTICS IN GAZEBO ENVIRONMENT."
Paper presented at 2018 International Mechanical Engineering Congress and Exposition (IMECE2018), Pittsburgh, PA, United States.
November 9, 2018 - November 15, 2018.

actual cameras. Gazebo is equipped with a camera sensor plugin which publishes images on a ROS topic. The overhead physical camera was emulated by introducing noise to the data streams generated by the virtual camera to help making the synthetic camera images more realistic because the Gazebo camera sensor model views the virtual world flawlessly. Gazebo also provides a sensor noise model which can add Gaussian noise parameters to the virtual camera sensor [19]. To achieve the goal of making our world experiment closer to the realistic environment, we implemented these Gaussian noise parameters and used three different values for noise standard deviation: σ = 0.007, 0.09 and 0.3.

Figure 12 shows images retrieved from the sensor with varying levels of Gaussian noise. The image on the left depicts standard deviation of σ = 0.003. In the central image, the virtual camera has standard deviation of σ = 0.03, which illustrates moderate level of noise. Finally, the image on the right represents the virtual camera with σ = 0.3 standard deviation, which shows high level of noise. A value of σ = 0.007 is considered to be reasonable for a decent digital camera [20].

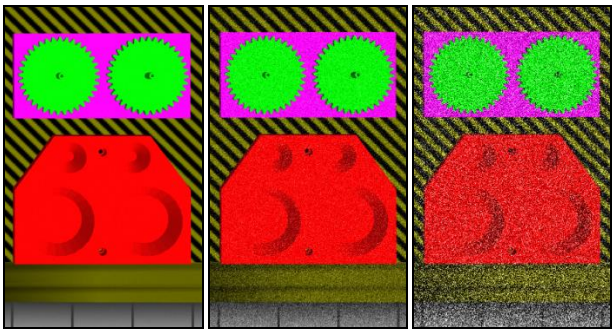


Figure 12. Image Data of the Virtual Camera Sensor with Low (on the left), Moderate (in the center), and High Level of Noise (on the right).

Sets of 1634 sequences of each of these three noisy camera streams were input to the object recognition system, which calculated three lists of the X-Y points. These are shown in Figure 13. Like Figure 11, these data showed normal-like distributions of the X and Y values, although the simulated images were much closer to a normal distribution and did not exhibit the clustering shown from real camera images. To more accurately replicate this clustering, the Gazebo noise model must be extended with a plug-in that allows for arbitrary customization, as noted in the description of future work in the concluding section.

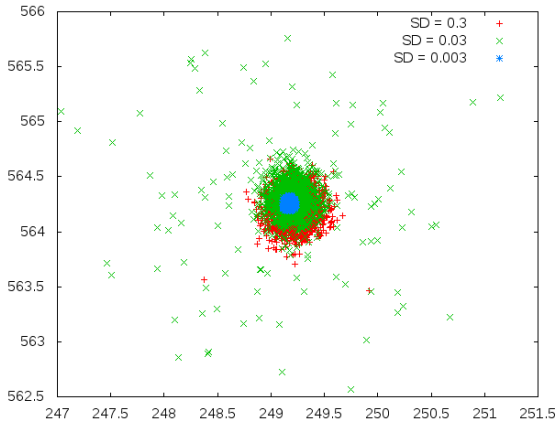


Figure 13. X-Y locations from the object recognition system for high (red), medium (green), and low (red) levels of simulated camera noise.

RESULTS AND CONCLUSIONS

This paper describes how the Gazebo physics-based simulation was applied to model the environment, objects, and robots in a laboratory used to measure the agility performance of robot systems. Simulation is an effective means that allows researchers to test their systems without tying up the actual robots, relieving the pressure on scarce resources and enabling safer and more repeatable testing. The goal was to replicate the noisy behavior of the real world in simulation, so that compensations and calibrations already built into the real planning and control system would function without modification. This necessitated performing effectively the inverses of the compensations and calibrations within the simulation. These apply to both the locations of the objects to be manipulated, and the poses of the simulated robots. Two methods of object pose adjustments were implemented, one that bypassed the object detection system and simply applied a noisy transform to the ideal poses, and one that included the object detection system and provided it with noisy camera images. Testing validated that the approaches achieved the intent, but could be improved with better registration of the simulated world with the physical world.

Toward the objective of testing strategies for sensor-based recovery from errors in a repeatable environment, the simulation is set up to allow either scripted configuration of initial testing conditions, or interactive placement of objects and robot starting locations. The integration of ROS with Gazebo allows for the timestamped logging of simulation states that can be compared between tests. Enabling hybrid real-virtual operation was achieved with minimal impact on the workcell configuration. CRCL interfaces and object detection reporting from the simulation followed the real robot protocols, with the selection of the real or virtual targets localized in a controller configuration that can be changed while the system is running. Another enhancement arose from the need to build a CRCL interface onto the Cartesian motion planner used with Gazebo.

Aksu, Murat; Michaloski, John; Proctor, Frederick.
"VIRTUAL EXPERIMENTAL INVESTIGATION FOR INDUSTRIAL ROBOTICS IN GAZEBO ENVIRONMENT."
Paper presented at 2018 International Mechanical Engineering Congress and Exposition (IMECE2018), Pittsburgh, PA, United States.
November 9, 2018 - November 15, 2018.

This provided another validation test of CRCL and its command-state protocol. While no changes to the CRCL XSD were required, the additional independent testing pointing to some assumptions on how to start and stop motions between pick-and-place activities, which were clarified in the documentation.

The overall research contribution is the use of simulation to supplement validation testing of a standard for knowledge representation of industrial robot tasks, CRCL. The use of simulation allows for more repeatable testing, enabled by the world model state inspection and logging features of Gazebo and ROS. A unique aspect of the research is the dual method of providing noisy sensor information to the object recognition system, using both noisy camera images and noisy object states, that attempt to match the clustering effects shown by the real system.

Future work includes doing better registration, including more capabilities of the real testbed such as shared tooling and vacuum gripping, and building simulated cameras that better match the actual cameras used. Ultimately, the simulation will be made available to external collaborators, such as those participating in the ARIAC challenges, to allow them to test their algorithms prior to running on the actual hardware in the APRS.

Disclaimer: Certain commercial/open source software, hardware, and tools are identified in this paper to explain our research. Such identification does not imply recommendation or endorsement by the authors or NIST, nor does it imply that items identified are necessarily the best available for the purpose.

REFERENCES

1. International Federation of Robotics, *Executive Summary: World Robotics 2017 Industrial Robots*. 2018: ifr.org.
2. Kootbally, Z., et al., *Enabling robot agility in manufacturing kitting applications.* Integrated Computer-Aided Engineering, 2018(Preprint): p. 1-20.
3. Koenig, N. and A. Howard. *Design and use paradigms for Gazebo, an open-source multi-robot simulator*. in *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*. 2004. IEEE.
4. Martinez, A. and E. Fernández, *Learning ROS for robotics programming*. 2013: Packt Publishing Ltd.
5. Agüero, C.E., et al., *Inside the virtual robotics challenge: Simulating real-time robotic disaster response.* IEEE Transactions on Automation Science and Engineering, 2015. **12**(2): p. 494-506.
6. Swanson, K.S., et al. *Extending driving simulator capabilities toward hardware-in-the-loop testbeds and remote vehicle interfaces*. in *Intelligent Vehicles Symposium Workshops (IV Workshops), 2013 IEEE*. 2013. IEEE.
7. Fernandes, L.C., et al. *Intelligent robotic car for autonomous navigation: Platform and system architecture*. in *Critical Embedded Systems (CBSEC), 2012 Second Brazilian Conference on*. 2012. IEEE.
8. Qian, W., et al. *Manipulation task simulation using ROS and Gazebo*. in *Robotics and Biomimetics (ROBIO), 2014 IEEE International Conference on*. 2014. IEEE.
9. Schlenoff, C., et al. *An IEEE standard ontology for robotics and automation*. in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. 2012. IEEE.
10. Proctor, F., et al., *The Canonical Robot Command Language (CRCL).* Industrial Robot: An International Journal, 2016. **43**(5): p. 495-502.
11. Fox, M. and D. Long, *PDDL2. 1: An extension to PDDL for expressing temporal planning domains.* Journal of artificial intelligence research, 2003.
12. Guimarães, W.H.P., et al., *Analysis of automated planning applied to an assembly and disassembly robot system.* 2013.
13. Mösenlechner, L. and M. Beetz. *Using Physics-and Sensor-based Simulation for High-Fidelity Temporal Projection of Realistic Robot Behavior*. in *ICAPS*. 2009.
14. Meeussen, W., J. Hsu, and R. Diankov, *URDF-Unified Robot Description Format*.
15. OSRF. *SDF*. 2014 [cited 2018 April 18]; Available from: http://sdformat.org/spec.
16. Diankov, R., *Openrave, ik fast module, openrave documentation*. 2016.
17. Proctor, F. *go motion*. [cited 2018 April 18]; Available from: https://github.com/frederickproctor/gomotion.
18. Horn, B.K., *Closed-form solution of absolute orientation using unit quaternions.* JOSA A, 1987. **4**(4): p. 629-642.
19. Newman, W.S., *A systematic approach to learning robot programming with ROS*. 2017, Boca Raton: Chapman & Hall/CRC ©2017. 530 pages.
20. OSRF. *Sensor Noise Model*. 2014 [cited 2018 April 24]; Available from: http://gazebosim.org/tutorials?tut=sensor_noise&cat=sensors.

Aksu, Murat; Michaloski, John; Proctor, Frederick.
"VIRTUAL EXPERIMENTAL INVESTIGATION FOR INDUSTRIAL ROBOTICS IN GAZEBO ENVIRONMENT."
Paper presented at 2018 International Mechanical Engineering Congress and Exposition (IMECE2018), Pittsburgh, PA, United States.
November 9, 2018 - November 15, 2018.

Benchmarking for Keyword Extraction Methodologies in Maintenance Work Orders

Thurston Sexton[1], Melinda Hodkiewicz[2], Michael P Brundage[1], Thomas Smoker[2]

[1] *National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, MD 20899*
thurston.sexton@nist.gov
michael.brundage@nist.gov

[2] *Faculty of Engineering and Mathematical Sciences, The University of Western Australia, 35 Stirling Hwy, Crawley WA 6009*
melinda.hodkiewicz@uwa.edu.au
thomas.smoker@research.uwa.edu.au

ABSTRACT

Maintenance has largely remained a human-knowledge centered activity, with the primary records of activity being text-based maintenance work orders (MWOs). However, the bulk of maintenance research does not currently attempt to quantify human knowledge, though this knowledge can be rich with useful contextual and system-level information. The underlying quality of data in MWOs often suffers from misspellings, domain-specific (or even workforce specific) jargon, and abbreviations, that prevent its immediate use in computer analyses. Therefore, approaches to making this data computable must translate unstructured text into a formal schema or system; i.e., perform a mapping from informal technical language to some computable format. *Keyword spotting* (or, *extraction*) has proven a valuable tool in reducing manual efforts while structuring data, by providing a systematic methodology to create computable knowledge. This technique searches for known vocabulary in a corpus and maps them to designed higher level concepts, shifting the primary effort away from structuring the MWOs themselves, toward creating a dictionary of domain specific terms and the knowledge that they represent. The presented work compares rules-based keyword extraction to data-driven tagging assistance, through quantitative and qualitative discussion of the key advantages and disadvantages. This will enable maintenance practitioners to select an appropriate approach to information encoding that provides needed functionality at minimal cost and effort.

1. INTRODUCTION

Maintenance is a vital function in every industry, including manufacturing, construction, chemical, infrastructure asset management, resource extraction industries, and many others. It involves all actions necessary to ensure a piece of equipment is in a state suitable to safely and consistently perform a required function (AS IEC 60300.3.14, 2005). Thus the related theory of maintenance practice can be split into *strategy development* and *work management* components. (Márquez, 2007; Kelly, 1997; Palmer, 1999) This paper focuses on the *work management* component of maintenance.

Work management includes the maintenance processes of work identification, planning, scheduling, execution, completion, and review. Data generated through these processes is typically captured using a maintenance work order (MWO), and while data about maintenance tasks differs from domain to domain (or even company to company within a domain), some or all of the following data are usually collected: 1) the asset and/or its components, 2) observed symptoms, 3) the time of failure, 4) the time for maintenance, 5) possible causes, 6) actions taken, and 7) the name of the technician(s). They often contain a mixture of human-generated, unstructured text, and structured field entries. These fields usually take the form of drop-down menus, lists, or entry fields, and times/dates for items such as order creation, progress, and completion.

This data is primarily recorded by the technicians who actually perform the maintenance; given that there are multiple technicians within an enterprise, the human-generated data is often inconsistent, error-filled, and replete with domain specific jargon. For example:

Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda; Smoker, Thomas.
"Benchmarking for keyword extraction methodologies in maintenance work orders."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

Technician A: *"bearing broken at Station 1"*
Technician B: *"bearing failure at cutoff unit of S1."*

Both of these representations describe the same overall problem of "broken bearing," located at "Station 1," but they take very different forms, especially if the end goal is to perform automated analysis (like finding all instances of "broken bearing" on this data). If this data could be parsed, it could lead to calculations such as failure mode identification, rework, problem spot identification, and more accurate mean time to repair (MTTR) or mean time between failure (MTBF), which can lead to improved maintenance strategy, reduced risk of failure and improved maintenance efficiency.

There have already been some successes in parsing these types of records in other domains. This is specifically true in the medical field, given the parallels between MWO records and patient medical records: both record symptoms, diagnoses and actions taken using unstructured text. Indeed, considerable work has been done on mining text in patient records; Heinze et al. (2001) applied natural language processing (NLP) across a wide variety of medical domains, while Tremblay et al. (2009) specifically applies it to fall-related injuries in veterans, and the *MedIE* system extracts and mines text from clinical medical records, generally (Zhou et al., 2006). While these efforts are significant achievements, the medical field has significant advantages over the domains we are concerned with:

1. data-sets tend to be much larger and cover longer time-spans by comparison, and
2. there are widely adopted controlled vocabularies available through medical ontologies.[1]

A number of research efforts exist in the engineering domains being addressed, that attempt to mitigate these issues.

1.1. Size of Datasets

How do maintenance practitioners in engineering obtain large quantities of data to robustly perform statistical analyses? In the authors experience within the manufacturing and mining equipment datasets, MWO dataset sizes for individual companies range from a few thousand records to upwards of one million MWOs each year. This quantity of MWO data is smaller than what most out-of-the-box solutions for NLP are built for (and regularly validated on), such as "tweets," or Amazon product reviews (Davidov et al., 2010).

One scheme to circumvent this issue promotes sharing data within a domain, and using a then-standardized dataset as training for data-driven tools that clean and analyze new data. This sharing is difficult when MWOs might contain proprietary information (machine identification numbers, technician names, specific processes for specific parts, etc.). There is ongoing research in anonymizing data-sets for this purpose, for example, focusing on "usefulness" as measured by information utility functions (Fang & Chang, 2008). Even using such an approach, information in a data-set will not be perfectly anonymized — there is a trade-off between privacy, and how much useful information remains for sharing.

1.2. Use of Ontologies

In the past, developers of Computerized Maintenance Management Systems (CMMS) have tried to ensure proper data structure through enforcing controlled vocabulary and problem code assignment for MWOs. In practice, these approaches have had limited success in improving data quality (Molina et al., 2013; Unsworth et al., 2011). With maintenance data especially, language used by one group (the maintenance technicians) can be quite different to that used by others (the engineers, or CMMS developers) (Murphy, 2010). Subsequently, the codes provided by engineers are often inadequate for expressing of the details of the event or action the maintainers take. Further, interpretations of events differ among the technicians themselves, and individuals might choose different codes for the same event.

There is growing interest in the potential value of ontologies to codify structures of *meaning* for maintenance. Early developments include the European project Proteus in 2005 from Rasoyska et al., with more recent work by, for example, Karray et al. (2012); Ebrahimipour & Yacout (2016). In the process-plant and engineering design sector, ISO15926 Standard Formal Ontology (ISO, 2003) could potentially be used for through-life support data. To date, however, there has been little uptake of ontological approaches by industry—in part because they have been developed in isolation. As a result, they are rarely interoperable, and lack scalability (Semy et al., 2004).

There remains a need for an agreed-upon upper ontology for maintenance. Current projects, such as the adaptation of Basic Formal Ontology (BFO) to the manufacturing sector (Arp et al., 2015), do include a sub-focus to provide an ontology for maintenance in manufacturing. Alternatively, the use of Natural Language Processing (NLP) to extract relevant information from the unstructured data sets promises to directly provide insights and analytics, even while maintainers continue to enter data in their own words. (Sharp et al., 2016; Sexton et al., 2017). This approach is somewhat ironically limited by the size of available training examples, mentioned previously.

[1] An ontology defines a machine-readable vocabulary to enable reasoning and with which queries and assertions are exchanged. Notable developments in medicine to underpin this capability include: SNOMED (Spackman et al., 1997), a nomenclature for human and veterinary medicine; the GENE Ontology for biology (Ashburner et al., 2000), a tool for the unification of biology; and the Unified Medical Language System (Bodenreider, 2004), a repository of biomedical vocabularies developed by the US National Library of Medicine.

Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda; Smoker, Thomas.
"Benchmarking for keyword extraction methodologies in maintenance work orders."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

1.3. Paper Outline

Informed by the dichotomy discussed, this paper compares two promising methods for automated data structuring through keyword extraction: a data-driven tagging method vs. a rules-based expert system. A publicly available mining equipment data-set is used to compare these methods for cognitive load on the human using these techniques, the ability of the method to calculate maintenance specific metrics (Median Time to Fail/ MTTF), and identification of problem-spots. The rest of the paper is structured as follows: Section 2 discusses background of both the rules-based method and the tagging method; Section 3 describes the data-set and how the two methods are compared, while Section 4 discusses these results; lastly, Section 5 presents conclusions and future work.

2. METHODS FOR ENCODING INFORMATION

We present a comparison of both previously discussed methodologies for encoding the tacit knowledge in MWOs into a more structured format. While obviously an incomplete overview of solutions to this common problem, we hope that the two selected methods are representative of two *archetypes* within the domain, namely: precisely-engineered, initially labour-intensive automation through design of "rules"; and data-driven, human-in-the-loop extraction that theoretically sacrifices precision for ease-of-use and statistical representativeness.

2.1. Rules-Based Methods

In rule-based data processing unstructured data is transformed into a predetermined format using explicit rule sets. These rule sets, often called "expert systems", are comprised of a series of 'if condition then perform action' statements where an action is performed if the given conditions are satisfied. Rule-based data processing requires progressive iteration of rule development and application in order to tune data sets to be capable of transforming unstructured data into the appropriate format (Rahm & Do, 2000; Prasad et al., 2011). An example output of a Rules-Based Method can be seen in Fig. 1.

It is important to have a purpose for the data structuring. In the case of maintenance work order records a common aim is to identify end-of-life events so that reliability metrics such as MTTF can be calculated. Other aims are to identify failure causes, track rework and develop troubleshooting capabilities. In each case a minimum viable data set to support the intended analysis needs to be identified. For calculation of time-to-end-of-life event, beyond just having a sufficient number of repeated events to sufficiently characterize the TTF distribution, the necessary data types are: an identifier of the asset or maintainable item that reached end-of-life, an identifier of the end-of-life event, the usage based on hours, distance, cycles or other measure to calculate life, and a means of identifying if the end-of-life event was due to censoring or not. Right censoring occurs when an item is removed before it reaches its end-of-life and when the observation period for data collection ends but the item is still in use.

Challenges in rule-based structuring include managing rule sets of increasing complexity and size. Rule sets are often executed sequentially with the order of rules important in determining the outcome of the transformation. Rule conflicts can occur which require the use of conflict resolution, often manual in nature to determine the appropriate output. As rule sets grow in size they become increasingly hard to manage, with each additional rule providing incrementally less benefit, yet with a possibility of degrading any previously executed rules.

2.2. Data-Driven Tags Method

Another approach to data-structuring is to derive patterns for recognition of good data using statistical aggregation, or any of several machine learning techniques. In this paradigm, text is processed in order to be represented "numerically". Previous work has compared several ways of using Natural Language Processing (NLP) on MWOs, including Bag of Words models and Word2Vec (Sharp et al., 2016). Regardless of the technique, the goal is to develop a computational representation of the text, that captures some fundamental statistics in the original language, and to develop a machine-learning (ML) pipeline to *predict* the correct organization of some set of work-orders. The primary issue then becomes creating a data-set to train the ML model, which is a labor-intensive task requiring at least a tacit rules system. Additionally, the amounts of data involved in this domain, as noted above, are smaller than typical use-cases for NLP, with high technicality, and not nearly sufficient examples to statistically cover the broad functionality-space.

To circumvent this issue, it is possible to use the concept of *tagging* as a form of user annotation of MWOs, to balance structure and flexibility. Tags are un-controlled, multi-label feature-assignments of text (or anything, potentially) that can be mapped quite easily to a bag-of-words representation. The problem now becomes creating a mapping between the existing language in historical MWOs and the set of tags that user might want to represent the MWOs through. Here, as in Sexton et al. (2017), we exploit statistical aggregation methods used in NLP (specifically, term-frequency/inverse-document-frequency weighting) to present users with the "most important" text-fragments—called "tokens"—*first*, allowing an annotator to generate a tag vocabulary list for post-facto automated extraction of these tags from historical MWOs. The output of *tagging* can be seen in Fig. 2. This allows the technicians to continue writing text with abbreviations and highly domain-specific, technical descriptions, while allow-

Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda; Smoker, Thomas.
"Benchmarking for keyword extraction methodologies in maintenance work orders."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

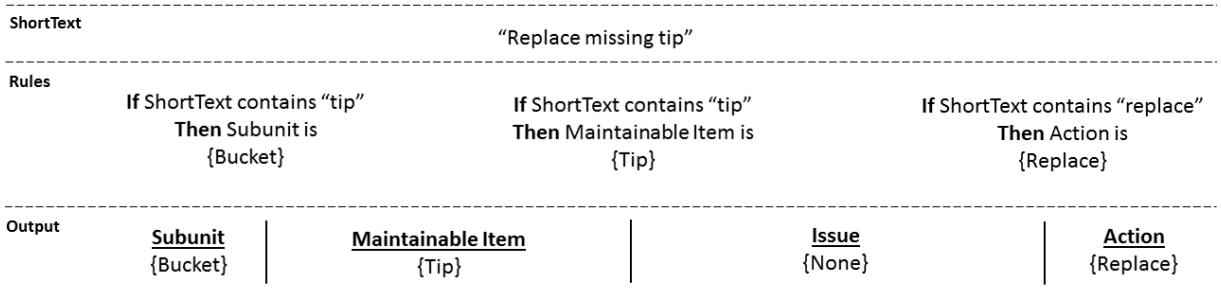


Figure 1. Illustrating the Rules Based Method for transforming MWOs. Rules are created given an expert's knowledge of the MWOs, however, they can grow with complexity as more situations are encountered. For example, if there are more "Maintainable items", the rules need to account for how to address multiple items. The Rules Based Method can handle misspellings, jargon, and abbreviations if the expert knows them beforehand. As seen in the figure, there is no rule for the "Issue" for this case, therefore, none is assigned.

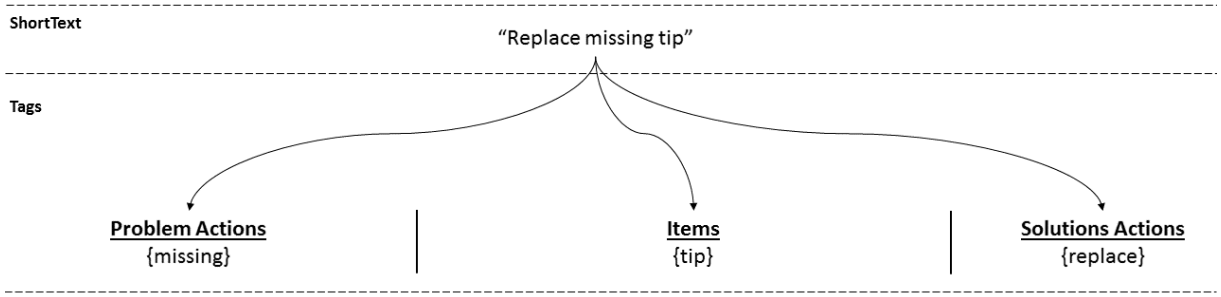


Figure 2. Illustrating the procedure for tagging a MWO. The tags are created using the original MWOs and are mapped to tags by a human expert. The output is the *Items*, *Problems*, and *Solutions*. This method creates structured data from the unstructured natural language MWO data. The tagging method can account for misspellings, jargon, and abbreviations in the original MWOs, but is dependent on how often these anomalies occur and how many tokens the human expert encounters. The tagging method, on its own, can not easily determine systems and subsystems (called Subunit in Fig. 1), while the Rules-Based Method can capture that information with the aid of a human expert.

ing an annotator/tagger to automatically extract tags from this text at their desired level of abstraction, thus drastically reducing the number of low-frequency concept-occurrences (see example experiment in Fig. 4).

Once a set of tags has been assigned to the set of MWOs, it is possible to perform queries by boolean set operations on tag occurrences. For example, if a sufficient number of related tokens have been mapped to the "broken" tag, a query over "broken" (conditional on particular machines) could be used as a good proxy for failure occurrence markers. Additionally, tag co-occurrences are naturally represented as graph "edges" between tag "nodes", making available a suite of graph database techniques and advantages.

3. DESCRIPTION OF EXPERIMENT

This comparative study consists of two steps: First, in the *information encoding* step, both methods described above are implemented on a single dataset, to perform the desired knowledge extraction. Subsequently, the structured forms are used to perform basic survival analysis, by approximating the labels for several major subsystems within the dataset, deriving these labels through the information encoded with both methods.

3.1. Information Encoding

The data set is an extract of maintenance work order records from a Computerized Maintenance Management database for five 1400 HP mining shovels. The data are publically available through the UWA Prognostics Data Library (Sikorska et al., 2016). Four fields are used in this analysis: *Date the Work Order is created, Asset identifier, Short Text,* and *Cost*. The *Short Text* field contains unstructured text populated by the individual generating the work order usually the asset operator, maintainers, maintenance planner or supervisor. Information that may be contained in the Short Text field is: the reason for the maintenance activity such as the observed symptoms of failure, a description of the work performed, the subunit(s) or maintainable items on which the work was performed and po-

Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda; Smoker, Thomas.
"Benchmarking for keyword extraction methodologies in maintenance work orders."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

sitional information (e.g right side, under). Typical examples of *Short Text* are 1) *Replace centre and LH lip shrouds* which describes the work performed (*(replace)* and the component *(centre and left hand lip shrouds)* and 2) *Broken grease line on bucket* which describes the problem *(broken grease line)* and the component (bucket). The data set used in this analysis (5485 records) has had the following cleaning prior to being made available in the data library: work orders were discarded due to issues such as incorrect functional location allocation, duplication, an absence of hours or costs logged and if the work order did not result in the repair or replacement i.e., inspections are not included.

3.1.1. Rules-Based Case Study

Rule files (also known as token files, not to be confused with NLP text "tokens") are constructed as a series of "if condition perform action"statements. Condition statements are comprised of between one to three logic statements: the location in the unstructured data-set to search; logical operator (choices of: equals, not equals, has, excludes, >, <); and patterns (regular expressions or numeric values) required to satisfy condition. Pattern matching (keyword spotting) is designed to be case insensitive and includes the ability to search through grammar, white space and alphanumeric search terms. Additional functionality includes: rule-conflict resolution to identify where multiple rules provide conflicting classifications on any given record, rule frequency statistics on how many times a rule action was activated, and records of the sequence of rules executed on each individual event.

Rules are developed using a piece-wise approach and stored in generic rule libraries, with one library per field to minimize interactions and mismatches between rule libraries. This allows the successful execution of other rules even in the presence of missing information that may cause other groups of rules to fail. The partitioning of rules allows development and reuse of rule libraries that can be used across multiple maintainable items. Generic rule libraries are developed by the selection and examination of training sets from more than one million MWOs for mobile mining assets such as haul trucks, shovels, excavators, loader and drills (Ho, 2015). These rule libraries are for a) maintenance action performed, b) failure mode or observed symptom of failure, c) maintainable item (for partial failures), d) location Identifier (e.g., "left rear" or "position 1"), e) active repair time, and f) down time. The development and subsequent reuse of generic rule libraries reduces development time.

All the rules are compiled into a token file. This token file was used to structure the data set for the five mining shovels used in this paper. The token file has 469 rules. Records require a minimum of three rules and can need as many as eleven.

Data about the problem and action are contained in the *Short-Text* field of the MWO. This has no set sentence structure, contains a high number of unique entries and technical jargon, abbreviations and spelling mistakes. Keyword spotting is used due to the prevalence of unambiguous keywords (e.g., "Engine" or "Leaking") which can be mapped directly to fields in the required minimum data-set such as asset or end of life event. Rules are developed manually to correct jargon, misspellings, variants and abbreviations.

The *Work-Order Type* and *Short-Text* fields are used to determine censoring status for the maintenance event. An event is classified as a failure if it contains one or more of the following criteria: *WorkOrder Type* code corresponding to a corrective maintenance code of the organization, a recorded failure mode, failure cause or symptom of failure in the *ShortText*, recording of a non-preventative maintenance action in the *ShortText*, recent job request or work order with a *ShortText* field recording symptoms of failure (job requests or work orders are classified as recent if they occurred since the previous scheduled service event), or active usage time of the maintainable item is less than half of the expected replacement interval recorded in the organization's maintenance plan. Events are classified as censored when the asset has been replaced; the work order type corresponds to the preventative maintenance code of the organization; and there are no recorded symptoms of failure or corrective maintenance actions, machine rebuilds or overhauls at fixed intervals, accidental damage, and secondary failures resulting from failure of a different maintainable items. Finally, all data is right censored at the end of the data collection period.

Data sets structured using the rule-based keyword spotting system are checked manually. Issues include duplicate rules and the need for conflict resolution if rule mismatches occur. For example when one rule classifies a work order as preventative yet another rule identifies the failure mode of breakdown. Identification of statistical outliers in time-to-failure data is used to review rules leading to the outliers. Other flags are when the time-to-failure interval is greater than fixed replacement interval specified by the organization's maintenance plan. Coded fields such as the Asset (also called Functional Location) and Work-Order Type fields can also contain inaccurate entries. Functional Locations fields are often recorded with values for an incorrect maintainable item or to a higher level in the functional location hierarchy. Numerical fields such as Total Actual Cost, Total Planned Costs and Man-hour fields are often populated by null values or with values that only reflect a partial cost of the maintenance work. Data quality issues in these coded and numerical fields necessitate the cross correlation of mismatching or missing information from text based fields. Interpretation of free-form text fields such as the Short-text field is required to cross-correlate with other data fields and extract relevant data elements that may be absent such as the identification of the maintainable item.

Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda; Smoker, Thomas.
"Benchmarking for keyword extraction methodologies in maintenance work orders."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

At the conclusion of data structuring an Excel file is created containing the new fields identifying the sub-system and maintainable item, the maintenance action, failure mode (if available), time since previous event, and a censoring indicator.

3.1.2. AI-Assisted Tagging Case Study

Starting from the same raw-text MWO descriptions as outlined above, text-fragment tokens are parsed from the entire corpus, and passed to a graphical user interface (GUI) that:

- presents an expert with tokens to "tag" in order of their TF-IDF score (Leskovec et al., 2014).
- suggests a list of potentially related tokens from the corpus, to promote terminology unification
- prompts a classification of each tag as "Problem", "Solution", or "Item".

An expert spent 60 minutes using this GUI to create and classify tags, saving their progress after every 10-minute interval. Fig. 3 demonstrates how quickly a large portion of the MWOs have a near-total classification rate—this can be calculated via *positive predictive value* (PPV)[2] by comparing the total number of raw tokens found in the work-order (i.e., the positives) to the number of tokens that have a valid tag mapping and classification (true-positives) [3] . Another way to measure the effectiveness of the tagging process as a form of terminology unification is by comparing the tag-frequency distribution to the original token frequencies (shown in Fig. 4). Far from the tokens—whose most common frequency is 1 by a large margin—tags created in this case-study are most likely to have between 5 - 15 occurrences, making this representation of the data much more amenable to statistical techniques.

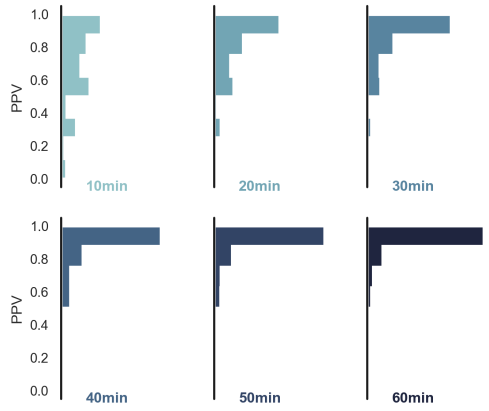


Figure 3. **Information encoding over time** — This figure demonstrates that the fraction of observed text-fragments/tokens that have a defined tag (with corresponding classification) increases rapidly as the tagger annotates the importance-ranked vocabulary list.

3.2. Application Case Study: Survival Analysis

To compare the usage of each structuring approach, a basic application of survival analysis is performed, with the goal of comparing the median time to fail for assets, according to several major subsystems. This can be done with both *parametric* and *non-parametric* models, from which Kaplan-Meier estimation and Weibull distribution models are represented here, respectively. All analysis was completed using the *lifelines* python package (Davidson-Pilon et al., 2018).

3.2.1. Kaplan-Meier Estimation

When there is a sufficient number of observations available, one can approximate the survival function of a population through a non-parametric estimator, the most well-known of

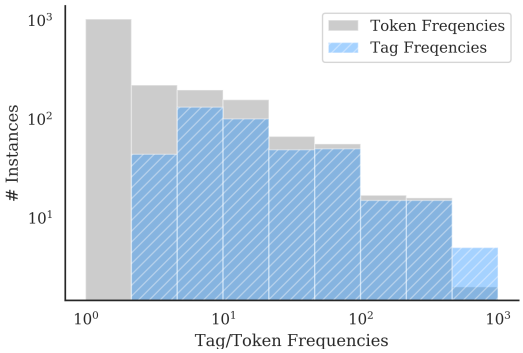


Figure 4. **Token vs. Tag frequency distributions** — the effect of mapping multiple low-occurrence tokens to some unified tag representation has a marked effect on the overall frequencies, dramatically decreasing the number of 1x or 2x occurring tags, and increasing the frequency of the most-recurring concepts, as is desired for statistical analysis.

[2] Typically called "precision" in an information-retrieval context.

[3] In theory, a complete rule-based method would map all observed, useful tokens to some structured information. In this way, we might assign the rules-based method a baseline PPV of 1.0 for all MWOs, to which the tag-based method is being compared.

Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda; Smoker, Thomas.
"Benchmarking for keyword extraction methodologies in maintenance work orders."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

which is the Kaplan-Meier (K-M) estimator (Meeker, 1998):

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \quad (1)$$

where each t_i is the time from initialization to the time of some event observation (a failure), d_i is the number of events occurring at t_i, and n_i is the number of population individuals known to survive or not have been censored at t_i. This allows censored data to be taken into account—for example, a part replaced in the course of scheduled maintenance has not *failed*, but is un-observable beyond its time of replacement, making this MWO a right censoring event.

Each point of $\hat{S}(t)$, then, is an estimate of the probability that any given member of a population will survive beyond the time t, given its previous survival up to that point.

3.2.2. Weibull Distribution Fit

It is often the case that, due to the rarity of actual failures for certain assets, one does not have enough data for a robust non-parametric estimate. In these cases, it is common among reliability engineers and others to assume that the set of "life-times" in a population is approximately exponentially distributed, and fit a distribution's parameters to available data. However, a strict exponential model assumes that the hazard rate of failure is constant in the population, meaning that the probability of failure is the same, no matter the age of an asset. When these properties of an exponential distribution cannot be assumed, then a Weibull distribution is often fit to the data; the survival function derived through this model (i.e., 1 − CDF) is given by:

$$S(t) = e^{-(t/\eta)^{\beta}}, \quad (2)$$

allowing for the β parameter to adjust the hazard rate as constant ($\beta = 1$), increasing ($\beta > 1$), or decreasing ($\beta < 1$). For this study, these Weibull parameters are estimated by fit the distribution to the observed failure inter-arrival times via Maximum Likelihood.

4. RESULTS & DISCUSSION

The primary way in which the two methods are compared is through the results of performing basic survival analysis on the data-set, after being structured with each approach. In the rules-based approach, each MWO is assigned to a "Major Subsystem" through application of one or more *rules*, along with a determination of whether the failure in this MWO was censored or not (through application of another rule). Since, in this data-set, the ID of each machine was noted in the MWOs, it is possible to calculate the running time for each machine between maintenance events, conditioned on each subsystem.

For the tag-based approach, it is necessary to approximate membership of each MWO into a subsystem by the set of tags extracted. The most straight-forward way to accomplish this is by selecting one tag that should be maximally representative of the subsystem (for example, the tag "bucket" for the bucket subsystem, etc.), and conditioning the failure inter-arrival times on that tag's occurrence. Obviously this will tend to under-estimate the number of failures, since there will be other objects or occurrences that are indicative of some particular subsystem. For example, if a boom is replaced, to which the bucket is attached (and therefore, also replaced), the "bucket" tag itself may not be explicitly extracted, since the bucket subsystem is only implicitly referenced via the "boom" tag. The single tag query for "bucket" would miss

Table 1. Results of the bench-marking experiment, organized by major subsystem. Queries for a set of multi-tag input $t_i \in T$ have an implied union: ($\cup T$). It is clear from the Weibull model that there are non-trivial decreases of the hazard rate occurring over time for all of the subsystems, but especially the "Engine" subsystem, and that this is indicated for both rules- and tag-based methods.

			MTTF (days)		Weibull Params.	
Major System	method	query	K-M	Weib.	β	η
Bucket	rules-based	Bucket	9.00	10.8	0.83±0.03	17±0.9
	single-tag	[bucket]	15.0	17.1	0.83±0.03	27±2
	multi-tag	[bucket, tooth, lip, pin]	9.00	10.5	0.82±0.02	16±0.9
Hydraulic System	rules-based	Hydraulic System	8.00	9.07	0.86±0.02	14±0.6
	single-tag	[hyd]	25.0	24.1	0.89±0.04	36±3
	multi-tag	[hyd, hose, pump, compressor]	9.00	9.74	0.89±0.02	15±0.7
Engine	rules-based	Engine	9.00	10.8	0.81±0.02	17±1
	single-tag	[engine]	10.0	11.8	0.79±0.03	19±1
	multi-tag	[engine, filter, fan]	8.00	9.31	0.81±0.02	15±0.8

Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda; Smoker, Thomas.
"Benchmarking for keyword extraction methodologies in maintenance work orders."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

this type of MWO. Additionally, the censoring of failure observations (here, the scheduled replacement of a part, e.g., before it had actually failed) was approximated with the occurrence of a "changeout" tag, for which the previous caveat also applies, but to action-words instead of subsystem item-words. The comparison between these single-tag estimates for Median Time to Failure (MTTF) and the rules-based estimates are shown in Table 1.

To approximate a remedy to the above "subsystem problem", and thus to derive more holistic approximations of the subsystem MTTF—with minimal annotation effort from a human—we also include an attempt by an expert to determine a reasonable set of subsystem-related tags, along with the corresponding approximation of the subsystem MTTF (to correspond to the rules-based method). We allowed the use of several (less than 5) tags that are *strictly members of the relevant subsystem*. These multi-tag approximations of the subsystem are simply the union of the set of MWO occurrences of each individual tag. As seen in Table 1, along with the K-M model comparisons shown in Fig. 5, these multi-tag approximations perform remarkably well at reaching a similar MTTF estimate to the rules-based methodology.

5. CONCLUSIONS & FUTURE WORK

In this study, two approaches to structuring unstructured data in the form of maintenance work orders were reviewed and bench-marked through the calculation of basic survival analysis models. While single-tag estimates tended to underestimate the failure rate, from Table 1, the average discrepancy between single-tag and rules-based estimates for MTTF, across the three tested subsystems and two methods, was only 7.7 days, with the majority of that discrepancy coming from the bucket subsystem (average of 16 days); when an expert is able to use his prior system knowledge, as was done through the previously-discussed multi-tag sets, that average discrepancy goes down to less than a day.

It is important to note that the methods discussed here are mainly compared between themselves — there is a distinct lack of a "gold-standard" measurement for, e.g., calculating the "true" subsystem MTTF, because the actual MTTF *per-subsystem* was never recorded in the first place. While it may be possible, going forward, to obtain such a well-curated reference dataset, the lack of this information speaks more broadly to the state of data availability and overall lack of standardized methodology through this process. We believe that the results here do not particularly advocate for one method over another; the rules-based keywords display a high level of thoroughness, but are only as complete as the number of hand-made rules being created, while the data-driven tags have a tendency to miss both rare events, and "obvious" physics-based relationships that inherently get encoded into a set of hand-made rules. Rather, we advocate for a combination of approaches going forward. The lack of a "gold-standard" is not uncommon in the broader information retrieval community, where the weighted opinions of "experts" are often combined to approximate an agreed-upon gold standard result. (Hripcsak & Rothschild, 2005)

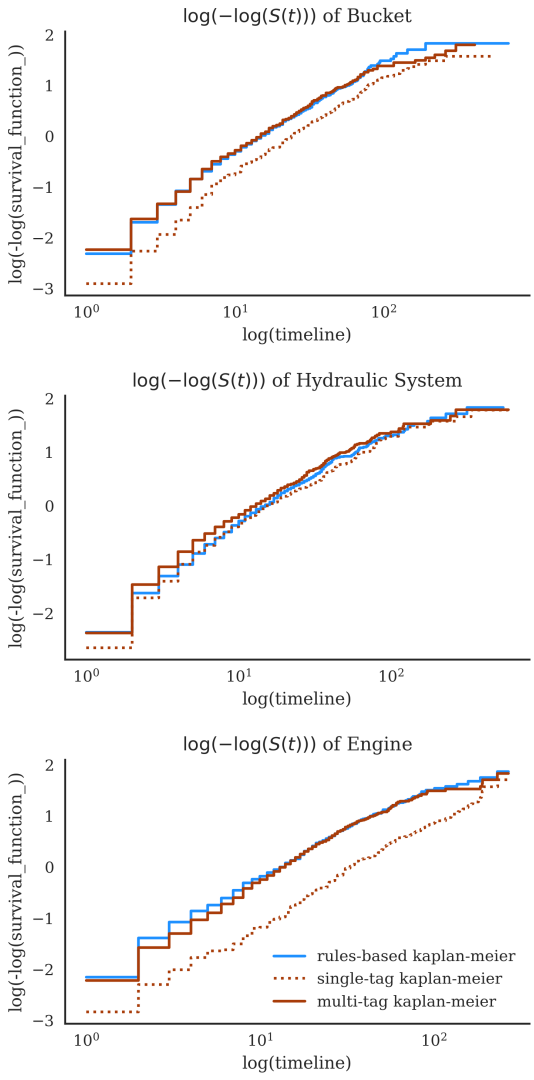


Figure 5. **Survival function comparison** — plotted on a log-scale, the multi-tag system approximation is clearly able to mirror the rules-based survival estimate across all relevant time-scales. Noticeable differences do occur for the lifespan extrema, though these effects are exaggerated in the plot and only last for small portions of the curve.

Given the at least one-week-difference in annotation labor time required between the two methods to achieve the reported results, the authors believe the tag-extraction method-

Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda; Smoker, Thomas.
"Benchmarking for keyword extraction methodologies in maintenance work orders."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

ology holds potential as an efficient tool for rapid MWO encoding. However, there are several key features of the rules-based approach that are lacking from the tag-based, and most important, perhaps, is the flexible definition of subsystem categorizations based on rule-matching. It would be very difficult to know, a priori, *which set of tags* and/or Boolean set operations would be "best" for approximating the classification of underlying subsystems.

Preferably, both rules-driven approaches, that encode some system-level from experts, *and* statistically sound empirical patterns from observation and data analytics, will continue to be explored as points of evidence toward a robust-yet-efficient standardized pipeline for encoding information from unstructured sources. Taking this further, we imagine a scheme where the development of taxonomies—or even ontologies—for expert systems are initialized and guided by latent patterns discovered from appropriate application of machine learning. Subsequent iterations of the machine learning pipelines for pattern discovery can then make use of human input via these "rule definitions", closing the loop that leads toward robust, hybridized, intelligence augmentation systems.

We suggest future efforts be directed toward the merging of automated tag extraction with the design of major functional relationships (encoded as rules), into an architecture for rapid, human-in-the-loop investigatory analysis. Such a system could take advantage of both the efficient data-processing from NLP techniques *and* the functional systems knowledge that human experts bring to the table.

ACKNOWLEDGEMENTS

This work was supported in part by the BHP Fellowship for Engineering for Remote Operations.

DISCLAIMER

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

REFERENCES

Arp, R., Smith, B., & Spear, A. D. (2015). *Building ontologies with basic formal ontology*. MIT Press.

AS IEC 60300.3.14. (2005). *Dependability management application guide - maintenance and maintenance support*. SAI.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... others (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, *25*(1), 25.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, *32*(suppl_1), D267–D270.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 107–116).

Davidson-Pilon, C., Kalderstam, J., Kuhn, B., Fiore-Gartland, A., Moneda, L., Zivich, P., ... Rendeiro, A. F. (2018, April). *Camdavidsonpilon/lifelines: v0.14.1*.

Ebrahimipour, V., & Yacout, S. (2016). *Ontology modeling in physical asset integrity management*. Springer.

Fang, C., & Chang, E.-C. (2008). Information leakage in optimal anonymized and diversified data. In *International workshop on information hiding* (pp. 30–44).

Heinze, D. T., Morsch, M. L., & Holbrook, J. (2001). Mining free-text medical records. In *Proceedings of the amia symposium* (p. 254).

Ho, M. (2015). *A shared reliability database for mobile mining equipment* (Unpublished doctoral dissertation). University of Western Australia.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, *12*(3), 296–298.

ISO. (2003). *Industrial automation systems and integration – integration of life-cycle data for process plants including oil and gas production facilities – part 2: Data model* (Tech. Rep.). Geneva, Switzerland.: International Standards Organisation.

Karray, M. H., Chebel-Morello, B., & Zerhouni, N. (2012). A formal ontology for industrial maintenance. *Applied Ontology*, *7*(3), 269–310.

Kelly, A. (1997). *Maintenance organization and systems*. Butterworth-Heinemann.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge university press.

Márquez, A. C. (2007). *The maintenance management framework: models and methods for complex systems maintenance*. Springer Science & Business Media.

Meeker, W. (1998). *Statistical methods for reliability data*. John Wiley Sons.

Molina, R., Unsworth, K., Hodkiewicz, M., & Adriasola, E. (2013). Are managerial pressure, technological control and intrinsic motivation effective in improving data quality? *Reliability Engineering & System Safety*, *119*, 26–34.

Murphy, G. D. (2010). Testing a tri-partite contingent model of engineering cultures: A pilot study. *Reliability Engineering & System Safety*, *95*(10), 1040–1049.

Palmer, D. (1999). *Maintenance planning and scheduling handbook*. McGraw-Hill Professional Publishing.

Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda; Smoker, Thomas.
"Benchmarking for keyword extraction methodologies in maintenance work orders."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

Prasad, K. H., Faruquie, T. A., Joshi, S., Chaturvedi, S., Subramaniam, L. V., & Mohania, M. (2011). Data cleansing techniques for large enterprise datasets. In *Srii global conference (srii), 2011 annual* (pp. 135–144).

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3–13.

Rasoyska, I., Chebel-Morello, B., & Zerhouni, N. (2005). Process of s-maintenance: decision support system for maintenance intervention. *Emerging Technologies and Factory Automation, 2005. ETFA 2005. 10th IEEE Conference on*, *2*, 8 pp.–686.

Semy, S. K., Pulvermacher, M. K., & Obrst, L. J. (2004). Toward the Use of an Upper Ontology for U . S . Government and U . S . Military Domains : An Evaluation Military Domains : An Evaluation.

Sexton, T., Brundage, M. P., Hoffman, M., & Morris, K. C. (2017). Hybrid datafication of maintenance logs from ai-assisted human tags. In *Big data (big data), 2017 ieee international conference on* (pp. 1769–1777).

Sharp, M. E., Sexton, T. B., & Brundage, M. P. (2016). Semi-autonomous labeling of unstructured maintenance log data for diagnostic root cause analysis.

Sikorska, J., Hodkiewicz, M., D'Cruz, A., Astfalck, L., & Keating, A. (2016). A collaborative data library for testing prognostic models. In B. Lung & B. Zhang (Eds.), *European conference of the prognostics and health management society 2016.*

Spackman, K. A., Campbell, K. E., & Côté, R. A. (1997). Snomed rt: a reference terminology for health care. In *Proceedings of the amia annual fall symposium* (p. 640).

Tremblay, M. C., Berndt, D. J., Luther, S. L., Foulis, P. R., & French, D. D. (2009). Identifying fall-related injuries: Text mining the electronic medical record. *Information Technology and Management*, *10*(4), 253.

Unsworth, K., Adriasola, E., Johnston-Billings, A., Dmitrieva, A., & Hodkiewicz, M. (2011). Goal hierarchy: Improving asset data quality by improving motivation. *Reliability Engineering & System Safety*, *96*(11), 1474–1481.

Zhou, X., Han, H., Chankai, I., Prestrud, A., & Brooks, A. (2006). Approaches to text mining for clinical medical records. In *Proceedings of the 2006 acm symposium on applied computing* (pp. 235–239).

Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda; Smoker, Thomas.
"Benchmarking for keyword extraction methodologies in maintenance work orders."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls

M. S. Hoehler[1] and B. Andres[2]

Abstract

This paper presents experimental investigations of the performance of common lateral force-resisting systems used in cold-formed steel construction under sequential thermal (fire) and mechanical (earthquake) loading. Wall specimens with gypsum-sheet steel composite sheathing, Oriented Strand Board (OSB) sheathing, or steel strap bracing were tested. The results demonstrate that the lateral capacity of wall systems can be reduced by exposure to fire. Additionally, fire performance of wall systems can be affected by pre-damage to the fire-resistive components that provide fire protection to these walls. The results are useful for fire compartmentation design when significant lateral deformation of a building is anticipated and post-fire assessment to repair or replace a structure. The study represents a step toward developing fire fragility functions for cold-formed steel framed shear wall systems to enable performance-based fire design.

Introduction

Although extensive information exists about the structural performance and fire resistance of cold-formed steel (CFS) construction; e.g. (Schafer et al. 2016; Sultan 1996; Takeda 2003; Wang et al. 2015), there is limited knowledge about the behavior of cold-formed steel lateral force-resisting systems (CFS-LFRS) under combined hazards; in particular earthquake and fire. In 2016, a series of experiments (Phase 1) was performed at the National Fire Research Laboratory at

[1]Research Structural Engineer, National Institute of Standards and Technology matthew.hoehler@nist.gov
[2]PhD Student, Danish Institute of Fire and Security Technology bav@dbi-net.dk

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

the National Institute of Standards and Technology (NIST) to investigate the performance of earthquake-damaged gypsum-sheet steel composite panel sheathed cold-formed steel shear walls under fire load (Hoehler et al. 2017). A second phase of the project (Phase 2) extends the study to two additional levels of fire severity and two additional types of CFS-LFRS: Oriented Strand Board (OSB) sheathed and strap braced walls.

The results provide data for a range of system performance under realistic fire conditions and can inform: fire compartmentation design when significant lateral deformation of the building is anticipated, post-fire assessment to repair or replace a structure, and first responder decisions to enter a building when earthquake aftershocks are likely. The study also represents a step toward developing fire fragility functions for cold-formed steel framed shear wall systems that will enable performance-based fire design of these structures.

Test Program and Specimens

Table 1 shows the Phase 2 test matrix. Three lateral force-resisting systems were investigated: gypsum-sheet steel composite panel sheathed walls, Oriented Strand Board (OSB) sheathed walls, and steel strap braced walls. The gypsum-sheet steel composite panels were a proprietary product where the gypsum was attached to the sheet steel by adhesive. The test specimens were subjected sequentially to combinations of mechanical (cyclic shear) deformation and thermal (fire) loading to investigate their post-fire lateral behavior as well as the sensitivities of the systems to pre-fire damage. Specimen names including ‘01’ were subjected only to load cycling to establish the baseline load-displacement response. Specimen names including ‘02’, ’03’, or ‘04’ were subjected to varied fire intensities followed by cyclic loading. Specimen names including ‘05’ or ’06’ were pre-damaged with cyclic loading, subjected to fire, and then cycled to failure. The influence of pre-damage on the performance of gypsum-sheet steel composite sheathed walls was investigated in Phase 1 (Hoehler et al. 2017). Specimens with an ‘R’ designation were either a test replicate or a redesign of the wall.

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

Table 1: Phase 2 test program

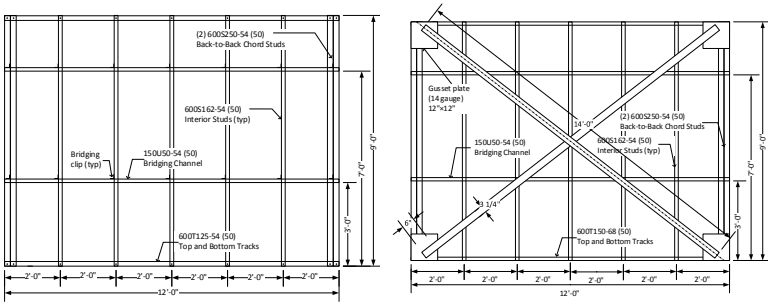
Wall Type	Specimen Name	Cycling (before fire)	Loading Fire	Cycling (after fire)
Gypsum-sheet steel composite	SB01	Cycle to failure	-	-
	SB02	-	Severe Parametric	Cycle to failure
	SB03	-	Mild Parametric	Cycle to failure
	SB04	-	ASTM E119 (1-hour)	Cycle to failure
Oriented Strand Board	OSB01	Cycle to failure	-	-
	OSB01R	Cycle to failure	-	-
	OSB02	-	Severe Parametric	Cycle to failure
	OSB03	-	Mild Parametric	Cycle to failure
	OSB03R	-	Mild Parametric	Cycle to failure
	OSB04	-	ASTM E119 (1-hour)	Cycle to failure
	OSB05	Drift Level 3	Mild Parametric	Cycle to failure
	OSB06	Drift Level 1	Mild Parametric	Cycle to failure
Strap braced	S01	Cycle to failure	-	-
	S01R	Cycle to failure	-	-
	S02	-	Severe Parametric	Cycle to failure
	S03	-	Mild Parametric	Cycle to failure
	S04	-	ASTM E119 (1-hour)	Cycle to failure
	S05	Drift Level 3	Mild Parametric	Cycle to failure
	S06	Drift Level 1	Mild Parametric	Cycle to failure
Additional	OSB01NG	Cycle to failure	-	-
	SB03R	-	Mild Parametric	Cycle to failure
	OSB_Kitchen	-	Real furnishings	-

Each of the specimens had a length of 12 ft. (3.66 m) and height of 9 ft. (2.74 m) and was designed using Allowable Stress Design nominally following American Iron and Steel Institute (AISI) standards (*AISI S400-15 w/S1-16, North American Standard for Seismic Design of Cold-Formed Steel Structural Systems (with Supplement 1)* 2016) and (*AISI S100-16 North American Specification for the Design of Cold-Formed Steel Structural Members* 2016). Both the gypsum-sheet steel composite panel sheathed walls and the OSB sheathed walls used the framing system in Fig. 1a. The framing system for the strap braced walls is shown in Fig. 1b. The cold-formed steel framing was 6 in. (150 mm) wide, had a specified strength of 50 ksi (345 MPa), and was connected using #10 screws (4.8 mm). #8 screws (4.2 mm) spaced at 4 in. (100 mm) along the panel edges were used to attach the gypsum-sheet steel composite and OSB sheathing. The strap braced walls were designed to achieve yielding of the steel straps.

All walls were designed to achieve a 1-hour fire-resistance rating per American Society for Testing and Materials (ASTM) standard (*ASTM E119-16a Standard Test Methods for Fire Tests of Building Construction and Materials* 2016). The

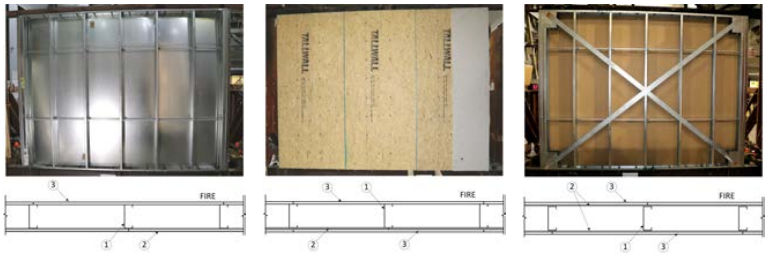
Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

cross sections are shown in Fig. 2. The design for fire-resistance of the gypsum-sheet steel composite panel sheathed walls was based on (*IAPMO-ER-1261 Sure-Board Series 200, 200W, and 200B Structural Panels Installed on Cold-Formed Steel or Wood Framed Shear Walls* 2018). The design for fire-resistance of the OSB walls was based on Underwriters Laboratory (UL) Design No. U423 (*UL Design No. 423 Fire Resistance Ratings - ANSI/UL 263* 2017) with the addition of wood panels as contemplated in (*Fire-resistance Ratings - ANSI/UL 263* 2017). The design for fire-resistance of the strap walls is based on UL Design No. U423. All walls used 5/8 in. (16 mm) thick Type X gypsum board with the joints taped and joints and fastener heads covered with one coat of joint compound on the fire-exposed side of the wall. The influence of insulation material in the wall cavity was not investigated.



(a) (b)

Fig. 1. Framing: (a) sheathed walls; (b) strap braced walls (1 ft. = 2.54 cm)



(a) (b) (c)

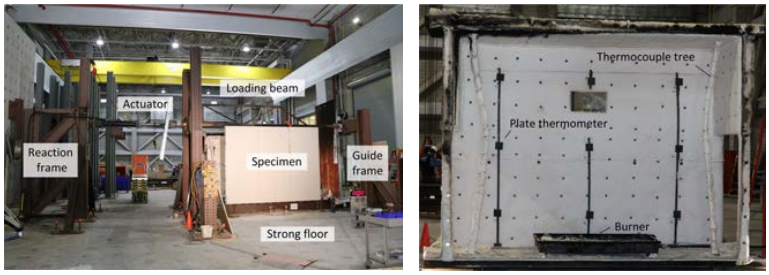
Fig. 2. Wall cross sections: (a) gypsum-sheet steel composite panel sheathed walls; (b) Oriented Strand Board sheathed walls; (c) strap braced walls (1 – steel framing; 2 – sheathing or straps; 3 – gypsum board)

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

Test Setup and Procedure

The test setup was informed by (*ASTM E2126-11 Standard Test Methods for Cyclic (Reversed) Load Test for Shear Resistance of Vertical Elements of the Lateral Force Resisting Systems for Buildings* 2011) but deviated as required to accommodate a burn compartment on a rolling platform. The test specimens were loaded mechanically by holding the base of the wall specimen fixed and applying a prescribed in-plane deformation to the top of the wall as shown in Fig. 3a. Out-of-plane movement of the wall was limited by four structural steel guide frames. Mechanical load was applied using a servo-hydraulically controlled actuator with a load capacity of 54 kips (240 kN) in tension and 82 kips (365 kN) in compression. Axial loading to the wall was limited to the self-weight of the specimen, actuator and top loading beam.

The thermal load was provided by a natural gas diffusion burner located in a movable compartment (interior dimensions: 9'-6" × 11'-6" × 4'-0" (2.9 m × 3.2 m × 1.2 m)). The constructed compartment is shown in Fig. 3b. The compartment was lined with two layers of 25 mm thick thermal ceramic blanket attached to sheet steel and cold-formed steel framing. The open side of the compartment that mated with the test specimen was lined with thermal ceramic blanket to provide a seal against smoke and flame leakage. The sides and top of the compartment overlapped the edges of the wall specimen approximately 3 in. (75 mm). The openings (vents) at the ends of the compartment were 5'-6" high by 4'-0" wide (1.4 m × 1.2 m).



(a) (b)

Fig. 3. Photographs of test setups: (a) mechanical loading; (b) fire loading

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

Mechanical Loading

ASTM E2126-11 Method C (CUREE Basic Loading Protocol) was used with reference deformations delta (Δ) of 1.5 % story drift for the sheathed walls and 2.5 % story drift for the strap braced walls. The loading procedure involves symmetric, reversed-cyclic displacement cycles grouped in phases at incrementally increasing displacement levels defined in the standard. The applied deformation was controlled using the actuator displacement. The displacement rates were selected to minimize inertial effects. With reference to Table 1, 'cycle to failure' was defined by a posted-peak load reduction of more than 70 % of the peak capacity. For tests with load cycling before the fire, for the sheathed walls 0.5 % and 1.5 % story drift were used for 'Drift Level 1' and 'Drift Level 3'[3], respectively. For the strap braced walls 0.5 % and 1.75 % story drift were used.

Fire Loading

It is assumed that the shear-resisting elements line a corridor and the fire occurs in a room adjacent to the corridor (Fig. 4). The target fire exposures were selected to represent various levels of fire severity. Three exposures were considered: (1) a 1-hour standard ASTM E119 fire curve, (2) a 'severe' fire exposure, and (3) a 'mild' fire exposure. The severe and mild fires represent realistic post-flashover compartment fire conditions with heating, fully-developed and decay phases. Fig. 5 plots the target temperature-time curves. The severity of the fire is defined in terms of exposure time and peak temperature. These values are informed by a statistical fit of data from compartment fire tests reported by (Hunt et al. 2010). Assuming a normal distribution of the compartment test data, 95 % of the reported peak compartment temperatures did not exceed 1100 °C and 50 % did not exceed 900 °C. These values were selected as the maximum temperatures for the 'severe' and 'mild' fires, respectively. Likewise, assuming a normal distribution of the duration of the fire, 70 min and 50 min represent 70 % and 50 % thresholds for the reported data, respectively. The length of the plateau was calculated using the time-to-burnout for the enclosure (τ_b) per (Hunt et al. 2010).

In multi-unit residential buildings, shear walls are commonly located along corridors adjacent to a kitchen. Assuming a kitchen compartment and taking the mean values of floor area and fuel load density reported by the National Research

[3] Intermediate 'Drift Level 2' was not investigated.

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

Council Canada (Bwalya et al. 2008) for multi-family dwellings (105 sq-ft (9.8 m^2) floor area with 805 MJ/m^2)[4], opening factors of 0.04 $m^{0.5}$ and 0.09 $m^{0.5}$ provide a time-to-burnout of 37 min and 16 min, respectively, using the Hunt et al. formulation. These times were rounded to 35 min and 15 min to define the temperature plateaus for the 'severe' and 'mild' fires. For comparison, the area under the target curve for the 'severe' fire represents a 20 % higher energy than ASTM E119 and the 'mild' fire corresponds to 40 % lower energy. The 'mild' fire is similar to the average upper gas layer time-temperature curves achieved in the Phase 1 tests.

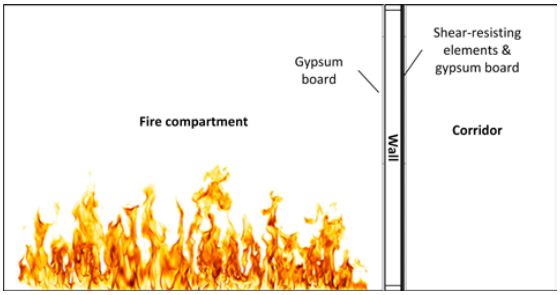


Fig. 4. Fire scenario for Phase 2 tests

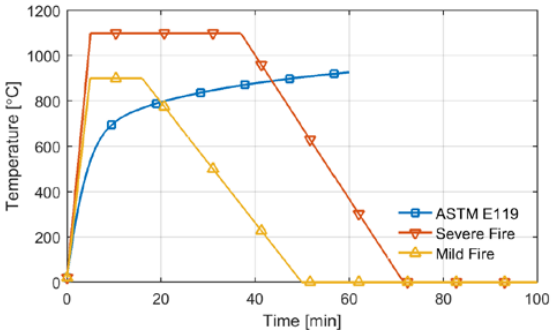


Fig. 5. Target upper layer gas temperature-time curves

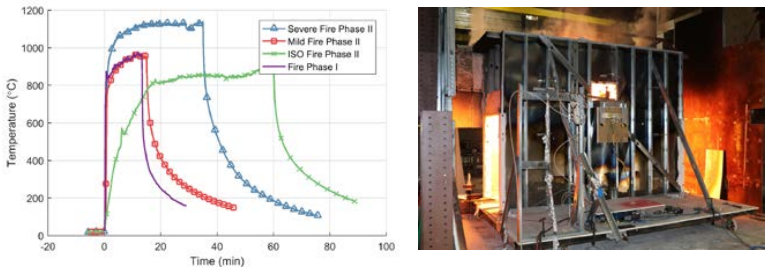
[4] Fire-related parameters are reported only in SI units because this is common practice in the U.S. and abroad.

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

Results

The experiments were completed immediately prior to the deadline for this paper. It is noted that the preliminary results presented here were collected from a limited series of experiments. Additional details and analysis will be included in future reports. The results presented here focus primarily on the structural, as opposed to thermal, behavior of the investigated wall systems.

The achieved upper layer gas temperatures in the compartment for the three fire scenarios investigated in Phase 2, as well as the comparable temperature measurement from Phase 1, are shown for the gypsum-sheet steel composite panel sheathed wall in Fig. 6a. The values are taken as the average of the top three sheathed, Chromel-Alumel thermocouple temperatures on the thermocouple trees at the north and south vents to the compartment (refer to Fig. 3b and Fig. 6b). The total expanded uncertainty (95 % confidence) for gas temperature measurement is estimated to be ±2.4 % of the reading. The compartment temperatures for the OSB walls exhibited greater variably in the ASTM E119 and severe fires due to the ignition of the combustible material in the wall. Fig. 6a emphasizes that the temperature rise for the mild and severe fires, which were based on simulations of real furnishing fires, appear more rapid than that in ASTM E119 test.



(a) (b)

Fig. 6. (a) measured average temperature of the three top thermocouples in both trees; (b) photograph of back of compartment during fire test

Fig. 7 shows photographs of the unexposed side (opposite to the fire compartment) of the walls during the severe fires where there was no pre-damage (cycling) before the fire. Fig. 8 shows the fire-exposed side of the walls after

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

cooling. The gypsum-sheet steel composite panel sheathed wall exhibited charring of the paper on the unexposed gypsum at the end of the heating phase (Fig. 7a), but the sheet steel remained in place (Fig. 8a) and kept flaming combustion inside of the compartment. The Oriented Strand Board in the OSB sheathed wall ignited during the heating phase (Fig. 7b) and was largely consumed during the fire (Fig. 8b). The gypsum opposite to the compartment in the strap braced walls was breached toward the end of the heating phase (Fig. 7c), but the straps remained in place through the cooling phase (Fig. 8c). Fire-induced oxidation of the straps on the upper south side of the wall (upper left in Fig. 8c) was observed. The damage to the wall by the ASTM E119 and mild fires was less severe and is illustrated using the post-fire load-displacement response of the walls in the subsequent plots. However, for all fire sizes, the gypsum on the fire-exposed side of the walls had lost almost all its strength after the wall had cooled, effectively preventing this layer of gypsum from contributing to the post-fire mechanical behavior of the wall.

(a) (b) (c)

Fig. 7. Unexposed side of wall during severe fire test: (a) gypsum-sheet steel composite panel sheathed wall 35 min after ignition (end of heating); (b) Oriented Strand Board sheathed wall 25 min after ignition; (c) strap braced wall 33 min after ignition (near end of heating)

(a) (b) (c)

Fig. 8. Fire-exposed side of wall after severe fire test: (a) gypsum-sheet steel composite panel sheathed wall; (b) Oriented Strand Board sheathed wall; (c) strap braced wall

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

Fig. 9 plots the applied actuator (lateral) load versus top-of-wall drift (measured on end opposite the actuator) during mechanical loading of gypsum-sheet steel composite sheathed walls. The total expanded uncertainty (95 % confidence) associated with the force and displacement measurements are 0.36 kips (1.6 kN) and 0.09 in. (2.3 mm), respectively. In this limited set of experiments, this wall system exhibited increasingly diminished post-fire capacity with increasing fire severity. The reduction in the peak load capacity was 23 %, 58 % and 68 % for the mild, ASTM E119 and severe fire, respectively. The mild fire effectively eliminated the gypsum on the fire-exposed side of the wall and partially degraded the adhesive on the composite panels (unexposed side) which allowed buckling of the sheet steel to occur. For information on the failure mode transition see (Hoehler et al. 2017; Hoehler and Smith 2016). The ASTM E119 fire further degraded the adhesive and more widespread buckling of the sheet steel occurred. In the severe fire, the fire oxidized (burned through) several screws along the top the wall and even burned through the sheet steel at a few locations. Nevertheless, the load redistributed and the system continued to resist lateral force.

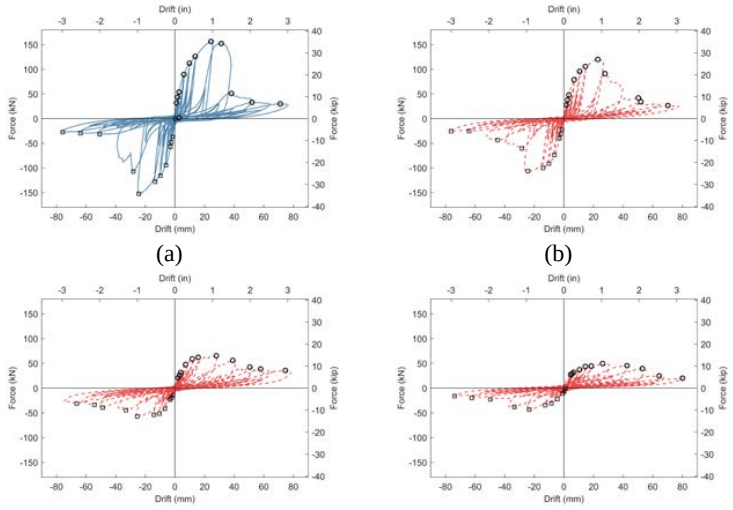


Fig. 9. Lateral load versus drift during mechanical loading of gypsum-sheet steel composite panel sheathed walls: (a) cycling without fire (SB01), (b) cycling after ‘mild’ fire (SB03), (c) cycling after ‘E119’ fire (SB04), (d) cycling after ‘severe’ fire (SB02)

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

Fig. 10 plots the lateral load versus drift during mechanical loading of OSB sheathed walls with no pre-damage prior to the fire. The investigated mild fire effectively eliminated the gypsum on the fire-exposed side of the wall and reduced the residual lateral capacity by 36 % (Fig. 10b). Both the ASTM E119 and severe fire caused the OSB to ignite. The burning was allowed to continue for 15 min after the burner was extinguished before it was suppressed with water. The reduction of the load capacity in both cases was nearly 100 % (Fig. 10c,d).

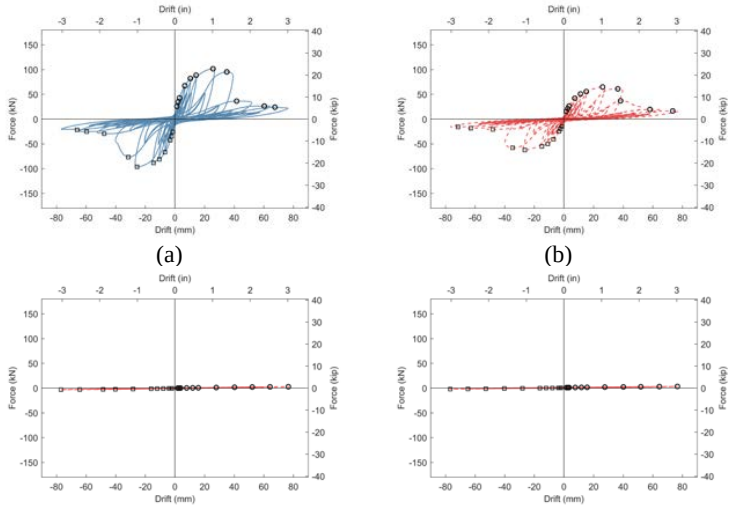
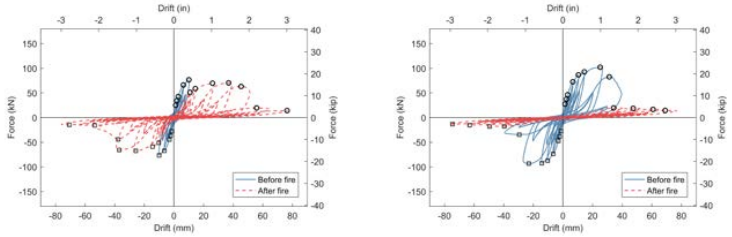


Fig. 10. Lateral load versus drift during mechanical loading of OSB sheathed walls: (a) cycling without fire (OSB01R); (b) cycling after ‘mild’ fire (OSB03); (c) cycling after ‘E119’ fire (OSB 04); (d) cycling after ‘severe’ fire (OSB02)

Cycling the wall to 0.5 % story drift prior to the fire resulted in minor damage to the skim coat on the gypsum board joints and no significant effect on the subsequent fire or post-fire cyclic performance; compare Fig. 10b to Fig. 11a. Cycling to 1.5 % story drift prior to the fire tore the tape along the joints and one of the OSB panels ignited during the mild fire. The fire was suppressed 15 min after the burner was extinguished. This burning degraded the post-fire capacity of the wall; compare Fig. 10b to Fig. 11b, however it is hard to see since the wall strength was already significantly degraded at 1.5 % drift.

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.



(a) (b)

Fig. 11. Lateral load versus drift during mechanical loading of OSB sheathed walls: (a) cycling to 0.5 % drift before 'mild' fire (OSB06); (b) cycling to 1.5 % drift before 'mild' fire (OSB05)

Fig. 12 plots the lateral load versus drift during mechanical loading of steel strap braced walls with no pre-damage (cycling) prior to the fire. The baseline hysteretic behavior Fig. 12a (ambient temperature) shows a pronounced peak near maximum load followed by a long plateau as the steel straps yielded. This peak is caused by the contribution of the gypsum boards on both sides of the wall. The failure mode was rupture of the straps at the gusset plate connections and/or crippling of the chord stud just above the hold-down at large lateral displacement (> 5 % story drift). The mild fire effectively eliminated the gypsum on the fire-exposed side of the wall and reduced the residual lateral capacity by 15 % (Fig. 12b). This reduction appears consistent with the loss of gypsum on the fire-exposed side of the wall. The response during the ASTM E119 fire was similar to that during the mild fire, however the gypsum paper on the inside of the wall on the unexposed side was blackened indicating higher wall temperatures. The reduction to the residual capacity (17 %) was similar to that during the mild fire (Fig. 12c). The severe fire burned through the gypsum on both sides of the wall toward the end of the heating phase (Fig. 7c). During subsequent cyclic loading, when cycling in the direction opposite to side where the oxidation of the straps occurred, the wall had almost zero residual load capacity (Fig. 12d, negative), while in the other loading direction close to the full ambient post-yielding load capacity was reached (Fig. 12d, positive). Interestingly, the post-fire ductility in this direction increase significantly (note axes scale change in Fig. 12d) and there was a more pronounced post-yielding hardening behavior for this limited set of tests. This appears consistent with the annealing of the cold-formed steel strap during the fire; but further study is required.

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

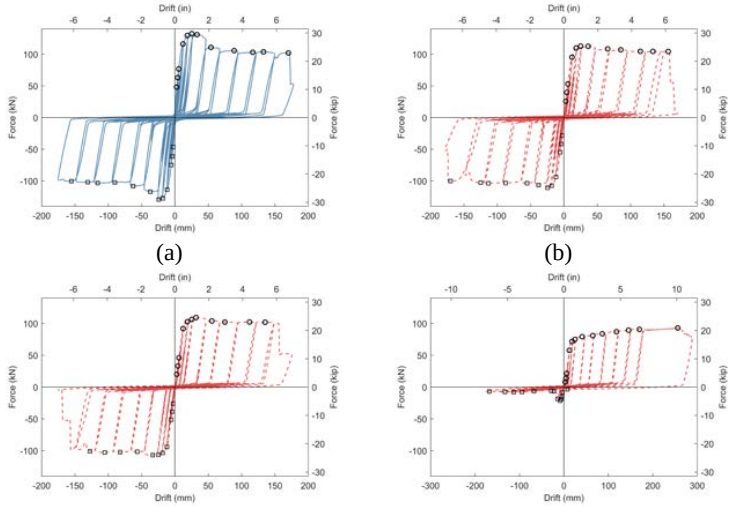


Fig. 12. Lateral load versus drift during mechanical loading of strap braced walls: (a) cycling without fire (S01R); (b) cycling after 'mild' fire (S03); (c) cycling after 'E119' fire (S04); (d) cycling after 'severe' fire (S02)

Cycling the wall to 0.5 % or 1.5 % story drift prior to the fire affected the contribution of the gypsum to the wall capacity, but had no discernable influence on the fire performance or post-fire yielding behavior (Fig. 13).

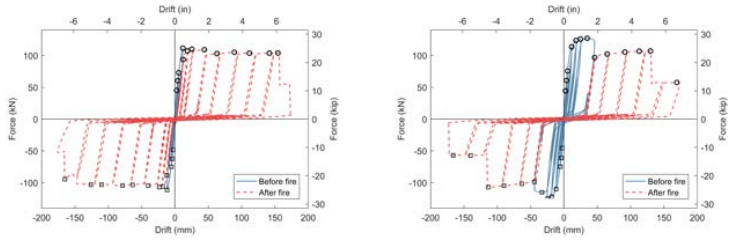


Fig. 13. Lateral load versus drift during mechanical loading of strap braced walls: (a) cycling to 0.5 % drift before 'mild' fire (S06); (b) cycling to 1.5 % drift before 'mild' fire (S05)

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

Conclusions

This research demonstrates an important interplay between the thermal (fire) and mechanical (cyclic) response of lateral force-resisting systems for cold-formed steel framed structures. The influence of a fire on the post-fire response differed significantly for the three investigated wall systems in this limited test series. The gypsum-sheet steel composite panel sheathing exhibited increasingly reduced post-fire capacity with increasing thermal assault. However, it maintained lateral load capacity in both loading directions even following the most severe fire investigated; allowing shear forces to redistribute even when some perimeter fasteners were burned away or the sheet steel had been comprised locally. The Phase 1 tests showed the composite panel system to be insensitive to cyclic damage prior to the fire. The strap braced walls were the most ductile and were largely insensitive to the thermal loading. However, in the case of the severe fire where a hotspot developed at a strap location, the residual lateral load capacity was reduced to essentially zero. The strap braced wall appeared to be insensitive to cyclic damage prior to the fire. For this limited set of experiments, the Oriented Strand Board (OSB) sheathed walls appeared to demonstrate a significant impact from the fire. Both the ASTM E119 and severe fires caused the gypsum-protected OSB to ignite, resulting in a total loss of residual capacity. Moreover, cycling to 1.5 % drift prior to the fire (as might occur in a major earthquake) allowed even the mild fire to penetrate the wall and ignite the OSB.

These are preliminary findings of a limited set of wall systems exposed to fire conditions. Analysis of this data is ongoing and additional testing is recommended. However, structural fire interactions such as those shown here have long gone uninvestigated and merit attention.

Acknowledgments

This work was funded by NIST. We thank Carleton Elliott (Sure-Board), Fernando Sesma (CEMCO), Jim DesLaurier (Marino/WARE), Brian Mucha (Panel Systems, Inc.), Larry Williams (SFIA), Benjamin Schafer (Johns Hopkins University), and Rob Madsen (Devco Engineering) for their expert consultation.

References

AISI S100-16 North American Specification for the Design of Cold-Formed Steel Structural Members. (2016). American Iron and Steel Institute

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

(AISI), Washington, DC.
AISI S400-15 w/S1-16, North American Standard for Seismic Design of Cold-Formed Steel Structural Systems (with Supplement 1). (2016). American Iron and Steel Institute (AISI), Washington, DC.
ASTM E119-16a Standard Test Methods for Fire Tests of Building Construction and Materials. (2016). ASTM International, West Conshohocken, PA.
ASTM E2126-11 Standard Test Methods for Cyclic (Reversed) Load Test for Shear Resistance of Vertical Elements of the Lateral Force Resisting Systems for Buildings. (2011). ASTM International, West Conshohocken, PA.
Bwalya, A. C., Lougheed, G. D., Kashef, A., and Saber, H. H. (2008). *Survey Results of Combustible Contents and Floor Areas in Multi-Family Dwellings*. National Research Council Canada.
Fire-resistance Ratings - ANSI/UL 263. (2017). Underwriters Laboratory (UL).
Hoehler, M. S., and Smith, C. M. (2016). *Influence of fire on the lateral load capacity of steel-sheathed cold-formed steel shear walls - report of test*. Gaithersburg, MD.
Hoehler, M. S., Smith, C. M., Hutchinson, T. C., Wang, X., Meacham, B. J., and Kamath, P. (2017). "Behavior of steel-sheathed shear walls subjected to seismic and fire loads." *Fire Safety Journal*, 91, 524–531.
Hunt, S. P., Cutonilli, J., and Hurley, M. (2010). *Evaluation of Enclosure Temperature Emperical Models*. Society of Fire Protection Engineers.
IAPMO-ER-1261 Sure-Board Series 200, 200W, and 200B Structural Panels Installed on Cold-Formed Steel or Wood Framed Shear Walls. (2018). International Association of Plumbing and Mechanical Officials (IAPMO).
Schafer, B. W., Ayhan, D., Leng, J., Liu, P., Padilla-Llano, D., Peterman, K. D., Stehman, M., Buonopane, S. G., Eatherton, M., Madsen, R., Manley, B., Moen, C. D., Nakata, N., Rogers, C., and Yu, C. (2016). "Seismic Response and Engineering of Cold-formed Steel Framed Buildings." *Structures*, 8, 197–212.
Sultan, M. A. (1996). "A Model for Predicting Heat Transfer through Noninsulated Unloaded Steel-Stud Gypsum Board Wall Assemblies Exposed to Fire." *Fire Technology*, 32(1), 239–257.
Takeda, H. (2003). "A model to predict fire resistance of non-load bearing wood-stud walls." *Fire and Materials*, 27(1), 19–39.
UL Design No. 423 Fire Resistance Ratings - ANSI/UL 263. (2017). .
Wang, X., Pantoli, E., Hutchinson, T. C., Restrepo, J. I., Wood, R. L., Hoehler, M. S., Grzesik, P., and Sesma, F. H. (2015). "Seismic Performance of Cold-Formed Steel Wall Systems in a Full-Scale Building." *Journal of Structural Engineering*, 141(10), 04015014.

Hoehler, Matthew; Andres Valiente, Blanca.
"Influence of Fire on the Shear Capacity of Cold-Formed Steel Framed Shear Walls."
Paper presented at Wei-Wen Yu International Specialty Conference on Cold-Formed Steel Structures 2018, St. Louis, MO, United States.
November 7, 2018 - November 8, 2018.

Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION

Sanjay Jain

George Washington University
Department of Decision Sciences
Funger Hall #415, 2201 G Street NW
Washington, DC 20052, USA

Anantha Narayanan

University of Maryland
Department of Mechanical Engineering
Glenn L. Martin Hall, 4298 Campus Dr.
College Park, MD 20742, USA

Yung-Tsun Tina Lee

National Institute of Standards and Technology
Engineering Laboratory
100 Bureau Drive
Gaithersburg, MD 20899, USA

ABSTRACT

Manufacturers need to quickly estimate cycle times for incoming orders for promising delivery dates. This can be achieved by using data analytics (DA) / machine learning (ML) approaches. Selecting the right DA/ML approach for an application is rather complex. Obtaining sufficient and right type of data for evaluating these approaches is a challenge. Simulation models can support this process by generating synthetic data. Simulation models can also be used to validate DA models by generating new data under varying conditions. This can help in the evaluation of alternative DA approaches across expected range of operational scenarios. This paper reports on use of simulation to select an approach to support the order promising function in manufacturing. Two DA approaches, Neural Networks and Gaussian Process Regression, are evaluated using data generated by a manufacturing simulation model. The applicability of the two approaches is discussed in the context of the selected application.

1 INTRODUCTION

Initiatives for advances in manufacturing, such as Smart Manufacturing, Industry 4.0, etc., call for exploiting the large volumes of data now available through ubiquitous sensors, to improve decision making by using Data Analytics (DA) and Machine Learning (ML) approaches. Though DA and ML are technically different, we use DA to cover both terms in this paper. The selection of the right DA approach is rather complex and requires significant expertise and effort. Manufacturing industry comprises of a wide range of configurations generally grouped into process and discrete production with multiple variations within each. Production of the same product by different companies may be set up using different philosophy. For example, a discrete product may be made using an assembly line setting, a batch production setting, or a job shop depending on the volumes and customization provided. Production systems may be configured differently even within the same production setting. For example, the division of work between sub-assembly lines and main line can vary for assembly lines for the same product. Beyond physical configuration, the operational policies can vary between push and pull, again with multiple further classifications within each. The level of technology employed to collect data and support real time decision making is another aspect with large variety. It is thus difficult to develop one approach that is applicable for manufacturing industry given the complexity due to the variety of product configurations, production system configurations, operation policies, and technology employed.

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

A number of DA approaches have been developed over the years, and new combinations continue to appear with ongoing research efforts. DA applications have been classified as descriptive, diagnostic, predictive, and prescriptive. DA approaches can fit multiple application classes. For example, Jain et al. (2017) discuss the use of simulation in diagnostic, predictive and prescriptive analysis roles and as a support application for other DA approaches. Recent efforts have focused on data-driven approaches to glean new learning from the data, and thus go beyond the modeling approaches based on known theoretical frameworks of systems. Such approaches include neural networks (NNs), Gaussian process regression (GPR), Bayesian networks, support vector machines (SVM), etc. It is rather complex to identify the approach that will work best for an application. In particular, the expertise required to identify the best approach for an application may not be generally available in manufacturing organizations.

This paper presents a simulation based approach for evaluating two DA approaches, NNs and GPRs, for a potential application in manufacturing. The two approaches were selected for this evaluation based on the reports in literature indicating these approaches generally worked better than other approaches (for example, see Scholz-Reiter et al. (2010)). The next section reviews relevant applications of NNs and GPRs in manufacturing and comparative evaluations of the two approaches in general. Section 3 discusses the use of simulation for generating the data for training and testing the two approaches. Section 4 describes our implementation of NN and GPR for predicting throughput in a small job shop environment, and Section 5 presents the experimental set-up used for this study. The results are presented and discussed in Section 6, and the last section concludes the paper.

2 RELATED WORK

This section identifies applications of NNs and GPR in manufacturing process and system control. Section 4 provides more technical details about NNs and GPR models and how we use them. For theoretical background on NNs and GPR, we refer the reader to Haykin (2004) and Rasmussen (2004). There are a number of applications of both approaches in maintenance area but they are not included due to space constraints. Comparative applications of the two approaches are included that are from outside of manufacturing due to lack of such works in manufacturing context.

2.1 Neural Networks for Manufacturing Applications

A number of applications of NNs have been proposed in the manufacturing domain at the process level with a few examples included here. Li et al. (2015) use a back propagation NN to optimize the cutting parameters in sculpted parts machining. The approach is shown to select process parameters leading to less machining time, less energy consumption, and better surface roughness compared to traditional approach for a test piece. Ding et al. (2016) use a NN for identifying the relationship between process parameters and aluminum bead geometry in arc-welding based additive manufacturing process. The authors use a Taguchi design for efficient data collection for training the NN. The weld settings are then selected based on the NN model. Khorasani and Yazdi (2017) similarly use a NN for identifying the relationship between milling process parameters and surface roughness of the product with a training data set collected using a full factorial design of experiment.

Recent reported applications of NN at levels higher than processes in the manufacturing management hierarchy are harder to find. Scholz-Reiter et al. (2010) cite an example of use of NN for dispatching rule selection from 2006 and a few other similar papers from the last millennium. Recently, Miller et al. (2014) train a NN for estimating the assembly times of vehicle sub-assemblies using geometric part information as inputs. The authors evaluate more than a hundred architectures of NN and use the top five to generate probability density plots of estimated assembly times. They achieved moderate success using this approach with predicted values falling within a +/-15% range of the associated target value. Rehman et al. (2016) develop a NN for estimating organization performance measures based on green manufacturing data on design initiatives, standards adaptation, purchasing, disposal, etc. for companies in the Indian industry. The

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

developed NN is applied to a steel company and was found to be useful in verifying company's initiatives and for providing further guidance.

2.2 Gaussian Process Regression for Manufacturing Applications

GPR applications in manufacturing focus at process level similar to NN applications and again a couple representative examples are provided here. Bhinge et al. (2014) use GPR methods to predict the energy used to machine a part based on machine monitoring data. They process the raw machine monitoring data into derived data that is then further process through a cutting simulator to generate operation parameters. GPR is applied with operation parameters as inputs and power consumption as output to develop an energy prediction model. Liu et al (2015) utilize an ensemble GPR approach to predict a measure of viscosity of the product of an industrial rubber mixing process based on parameters and recipes. The authors also include the ability of GPR based approach to generate uncertainty measures for the predictions as an advantage over other available approaches for the purpose.

2.3 Comparison of NNs and GPR

There are a handful of efforts available in literature that compare NNs and GPR in application settings. Goebel et al. (2008) compared three DA approaches, NNs, GPR, and Relevance Vector Machines (RVM), for prediction of remaining useful life based on damage for a rotating aerospace equipment on test stand. The set up was limited to a small damage data set given the expensive set up. The authors found NNs performance was dependent on choice of data and the design of its architecture. Performance of GPR was similarly dependent of the choice of the covariance function used but it offered the advantage of providing confidence bounds around mean predictions. GPR was identified as scaling typically as $O(n^3)$ with the number of training data points and thus would demand high computation power and time. The authors also indicated the need for domain specific measures to compare performance beyond accuracy.

Scholz-Reiter et al. (2010) compare GPR with NN and other approaches for dynamically selecting dispatching rules in production scheduling. They train the approaches using utilization and due date tightness as inputs and tardiness as outputs. They found that GPR generally outperformed other approaches including NN. Interestingly, their results differ from all other comparative studies as they indicate that GPR is outperformed by NN for smaller learning data sets. They identify some issue with hyperparameter setting of GPR for this anomaly and propose to further investigate it.

Ahmed et al. (2010) compared eight machine learning approaches including multiple variations of NNs and GPR for business-type time series. They used monthly time series of the benchmark M3 competition data (IIF 2017) for the study. The multilayer perceptron (MLP), referred to as NN in this paper, and GPR were found to be the top performing approaches. The performance was found dependent on the preprocessing of the data with MLP performing best for the commonly used lagged-value and moving average techniques. GPR was found to have second best performance with the two preprocessing techniques but was found to be robust for the difference technique also.

Chen et al. (2014) compare GPR with NN for wind power forecasting and find that GPR provided 9-14% improvement over NN for two large data sets. They note the advantage of GPR improved to 17% for the third data set with limited amount of training data.

Kamath et al. (2018) compare the two for representing potential energy surfaces in molecules with more than 3 atoms. They found the GPR outperformed NNs, that is, achieved higher accuracy with smaller data sets used for fitting. However, they point out that GPR is slower compared to NNs and needs training data points that are sufficiently far apart. NNs can work with overlapping data points, but can suffer from overfitting. NNs were considered easier to build and recommended when the cost of data collection is low. The authors suggested that both approaches can gain from optimized sampling of the training data points.

The comparative studies report above do draw some common conclusions across the widely different application areas. GPR appears to have an accuracy advantage over NNs when training data size is limited. Also, GPR is identified as computationally expensive approach compared to NNs. Only one comparative

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

study by Scholz-Reiter et al. (2010) mentioned above used a manufacturing application scenario and interestingly it differed in its assessment of GPR performance for smaller data sets from all the other studies cited above. The different results suggest that the performance of the DA approaches may depend on the application. With the push towards smart manufacturing there is a need to evaluate the potential manufacturing system level planning and control applications of DA approaches.

3 DATA GENERATION USING SIMULATION

DA approaches such as NN and GPR require data for training the respective models that can be used in a predictive capacity. While the models can provide predicted outcomes for input data for situations that were out of the coverage range in the training data, such predictions may have low accuracy. It is best to have such predictive models be trained across the range of situations that they will then be required to assess. For manufacturing applications, it can take a long time and lot of effort to collect such data from the real manufacturing system. Simulation models of manufacturing allow the option of creating a range of situations in the virtual representation and collecting the necessary data for training the predictive models.

Also, it is generally difficult to access data from manufacturing organizations for researchers. This paper, hence, uses a virtual factory prototype, essentially a multi-resolution simulation model of manufacturing, for generating the data. This section provide brief information on the use case, the virtual factory prototype and the set-up for data generation. The reader is referred to Jain et al. (2017) for more details.

3.1 Use Case

The use case for this work is based on the order promising scenario for a small job shop that produces three part types that are essentially the same part but produced using different materials: aluminum, titanium, and steel. The small job shop consists of a turning cell with 4 machines and a milling cell with 2 machines. All parts go through two operations, the first one being in the turning cell and the second one in the milling cell. Each arriving batch goes from a raw material stock to a finished good area after being successively processed in the two cells. Each batch is composed of ten parts. Each type of part requires different value ranges of machine parameters leading to different processing times. Orders are received at the source and specify the part type and the number of parts to be processed in the shop. In the base scenario, the order frequency follows a normal distribution with a mean of 60 minutes. The order frequency is varied to mimic different load levels in the shop.

Customers place orders for defined quantities of one of the three part types. The planner performing the order promising function has to provide an expected shipment date for the order. It is complex to estimate the shipment date as it can vary based on several factors including the ordered quantity, current load on the shop, and machine failure characteristics. Simulation models that use the current situation on the shop floor as the starting condition can be used for estimating the shipment date, but it can take time to generate such estimates and may require more expertise than a typical planner may have. Also, the company wouldn't want to have customers wait for several minutes for finding out the shipment date. The planner needs to be supported by an application that generates shipment date estimates for proposed orders within seconds. The two DA approaches are being evaluated to meet the need for estimating the cycle time for an incoming order based on the current conditions on the shop floor.

3.2 Virtual Factory Prototype

The virtual factory prototype is an initial effort towards implementing the virtual factory concept described in Jain et al. (2017). The prototype allows development of integrated multi-resolution model that can be executed at the selected resolution. The levels include a physics-based process representation, an agent simulation based machine representation, and a discrete event simulation based cell or factory representation. It includes the capability of reading in data files describing a small manufacturing system,

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

generating a discrete event simulation model using a library of machine level models, executing the model together with a basic animation, and generating selected output graphs.

3.3 Data Generation Set-up

The data generation set-up is shown in Figure 1. The virtual factory model has been set-up with standards based interfaces. The input files include manufacturing configuration data in Core Manufacturing Simulation Data (CMSD) standard format (SISO 2012), machine instructions in STEP-NC format (ISO 2007), and some custom formats for machine data. The virtual factory model is generated mostly automatically using the input data files together with the library. Following the execution of the model, in addition to the standard results generated by the simulation software, output files are generated using standard formats as shown in the figure. The factory data files in Business To Manufacturing Markup Language (B2MML; MESA 2013) are used for training the DA approaches for this study.

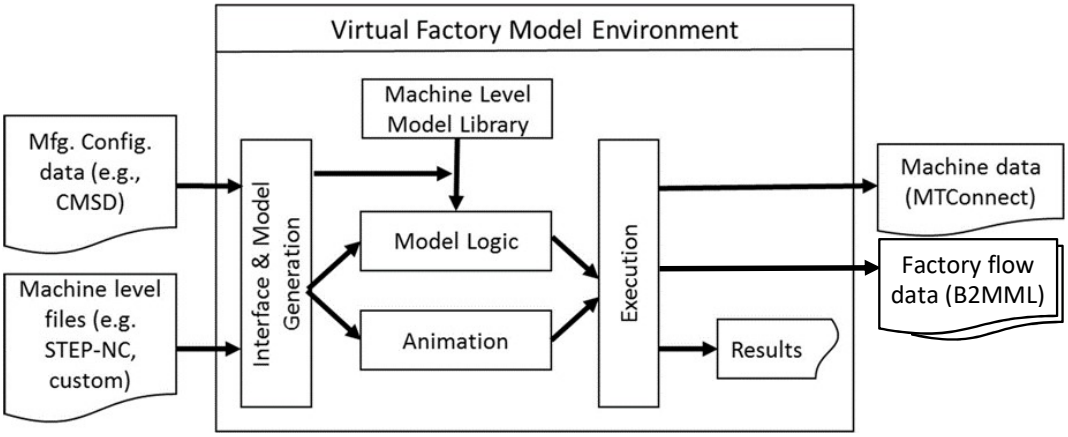


Figure 1: Data generation using virtual factory prototype.

4 METHOD IMPLEMENTATIONS

4.1 Neural Network Implementation

The NN implementation was described in Jain et al. (2017) and is briefly summarized here. Artificial Neural Networks (ANNs) are a type of predictive model that consist of an input layer, a number of hidden layers, and an output layer. Each layer consists of one or more neurons, and the neurons between layers are connected by weighted edges. The neurons of the input layer correspond to the known system variables, and the neuron in the output layer corresponds to the metric to be predicted. The ANN consists of a series of linear and non-linear transformations that turn the input variables to the output metric. During training, the weights on the edges of the NN are adjusted to produce the correct output value for each combination of input values from a training data set.

In this work, the output value to be predicted is the duration expected to complete an incoming job. The input values are the number of parts in the job, the material type, and the current load on the system. The current load on the system is modeled by a triplet (n_A, n_S, n_T), which denotes the parts of material type aluminum, steel, and titanium currently being processed in the system. The NN had one hidden layer with ten neurons. We generated a large data set from simulations. The data set was split into two parts, one for training the NN, and the other for validating the trained NN model. The details of the NN implementation and the results of the validation are described in detail in Jain et al. (2017).

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

4.2 Gaussian Process Regression Implementation

Gaussian process regression (GPR) is a probabilistic method of interpolation to determine a target value from given inputs. Instead of computing a single polynomial with a fixed number of parameters to fit the training data, GPR determines a distribution of random functions that best fit the data. The distribution of random functions that fits the data (called the posterior) is determined from a prior distribution of random functions, which is defined by a covariance and mean. GPR uses a Gaussian distribution for the priors.

The goal of GPR is to determine an unknown target function *f(x)* from the prior distribution and some known data points. To define the prior distribution, we use a kernel function that approximates the covariance, which is a measure of the geometrical distance of closely located input points and their corresponding function values. The chosen kernel function determines the geometrical shape of the target function, and depends on the scenario and data we are interested in. There are many choices for the kernel function. In this study, we chose the radial basis function (RBF), also called the squared exponential, which is a commonly used in many situations.

The squared exponential covariance function is defined as below:

$$K(x, x') = \sigma^2 exp\left[-\frac{1}{2}\left(\frac{x - x'}{l}\right)^2\right] \quad (1)$$

where, $K(x, x')$ represents the covariance function for the pair of inputs x and x', and σ and l are the hyperparameters that represent the amplitude and length scale, respectively. The amplitude signifies the overall magnitude of the covariance value, and the length scale indicates the relevance of the input features to the response y. In simple terms, adjusting σ changes the overall magnitude of the covariance, and adjusting l changes the smoothness of the target function curve. These hyper parameters must be tuned appropriately to obtain a good target function that matches the data.

5 EXPERIMENTAL SET-UP

The data generation using the virtual factory prototype is described in Jain et al. (2017), and summarized briefly in Section 5.1. Section 5.2 describes the data generated for training and validation for this study.

5.1 Data Generated using the Virtual Factory Prototype

The virtual factory prototype was used to mostly auto-generate a model of the small job shop described in Sections 3.2 and 3.3. The execution was set up to run primarily at the cell level of detail using the discrete event simulation model in the hierarchy of models generated in the prototype for the scenario. Corresponding to the execution at the cell level of detail, the model was set up to generate the factory flow data files in B2MML format (MESA 2013) while the generation of machine event data stream was disabled. In the simulated scenario, raw material is processed through a turning cell, and then a milling cell, and then ends in the finish goods area. The simulation of failure and repair time of machines is included. The time to failure follows an exponential distribution with a mean time between failures (MTBF) defined in the virtual factory model. The machine model remains in the failure state for a sampled value of repair time. The repair time follows an exponential distribution with a mean time to repair (MTTR) also defined in the virtual factory model.

Two different B2MML files were generated. First file recorded the total number of parts produced for each part type and the total duration to produce those parts. The second file recorded the following information for each order completed: ID, start and end times, part type and number of parts ordered, and the load of the factory at the time the order was released captured as number of parts of each of the three part types in process in the job shop at the time. The B2MML files are converted to comma-separated value (CSV) format to facilitate further processing by the DA approaches.

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

6 RESULTS AND DISCUSSIONS

In this section, we discuss the results of the tests. First we look at the overall prediction by the GPR model, and compare it with the performance of the NN. Figure 2 shows the graph of the predictions for the test data set. The predicted values from the GPR model are represented by a blue line with diamond markers, the predictions of the NN model are represented by a green line with '+' markers, and the corresponding test ground truth values are represented by an orange line with dot markers. The gray area represents the confidence bounds of the GPR prediction, up to one standard deviation above and below the mean. Note that the *x*-axis is simply a sequence of incoming orders – and does not show the other parameters associated with that order. Each point on the *x*-axis corresponds to a point in the test data set that includes the following values: the number of parts in that order, the material type for that order, and the current load on the factory denoted as a triplet (n_A, n_S, n_T). It is expected that in a realistic system, the parameter representing load on the system will be a continuously monitored variable of the system, rather than a part of the order specification. For the purpose of our study, we denote this parameter explicitly as an input parameter.

In the case of the GPR model, the root mean squared error (RMSE) was 1626 seconds, and the mean absolute error (MAE) was 1241 seconds for the test data set. The hyperparameters of the model were empirically chosen. The mean duration in the test data set was 61426 seconds. This represents an error of about 2% on average in the predicted order-completion duration. The NN model had a poorer performance, with an RMSE of 4013 and MAE of 2414. As can be seen in Figure 2, the NN model performance suffers in the range of orders between Order Number 200 and 500, which corresponds to a high load on the system in the test set. The GPR performs better in this region, and not only produces more accurate average predictions, but also provides confidence bounds that reflect the confidence in the prediction, which can

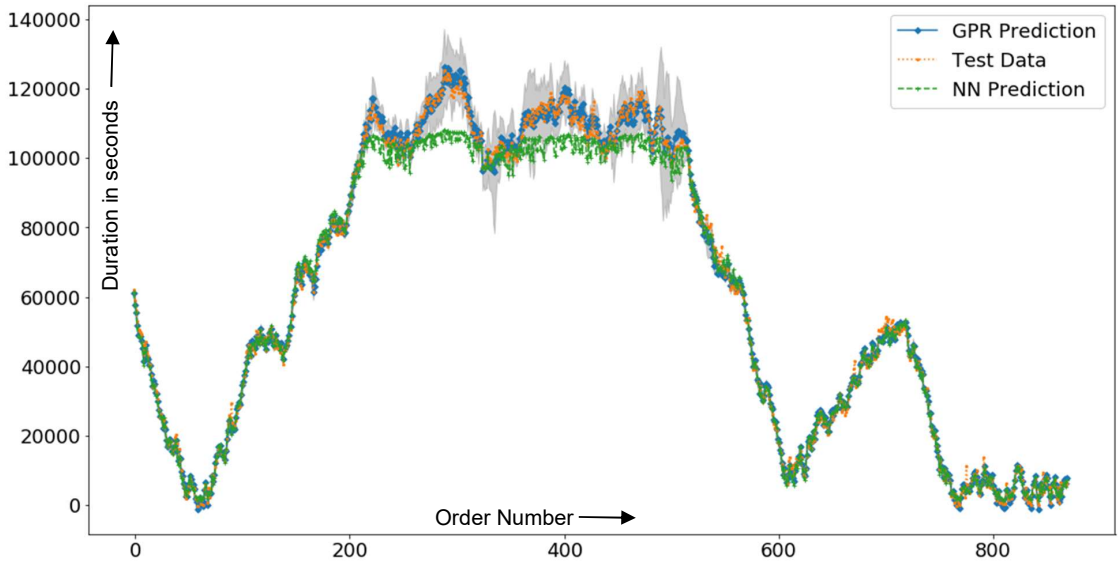


Figure 2: Order-completion duration predictions for test data.

help decision making.

It may be possible to tune the models to further reduce the errors. From the test results, it can be observed that the predictions of both models are less accurate in the range of orders corresponding to the factory being under heavy load in the simulation, as seen by the orders that require a long duration to complete from the moment of order arrival, due to the number of jobs already running on the machines and waiting in the queues. For clarity and comparison, Figures 3 and 4 show expanded views of different portions of the same test data in Figure 2. Notice the difference in order duration on the *y*-axis of the two

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

figures. For the high duration orders in Figure 4, it can be seen that the predictions are less accurate. Some additional model tuning may be performed to address this portion, and is left as future work.

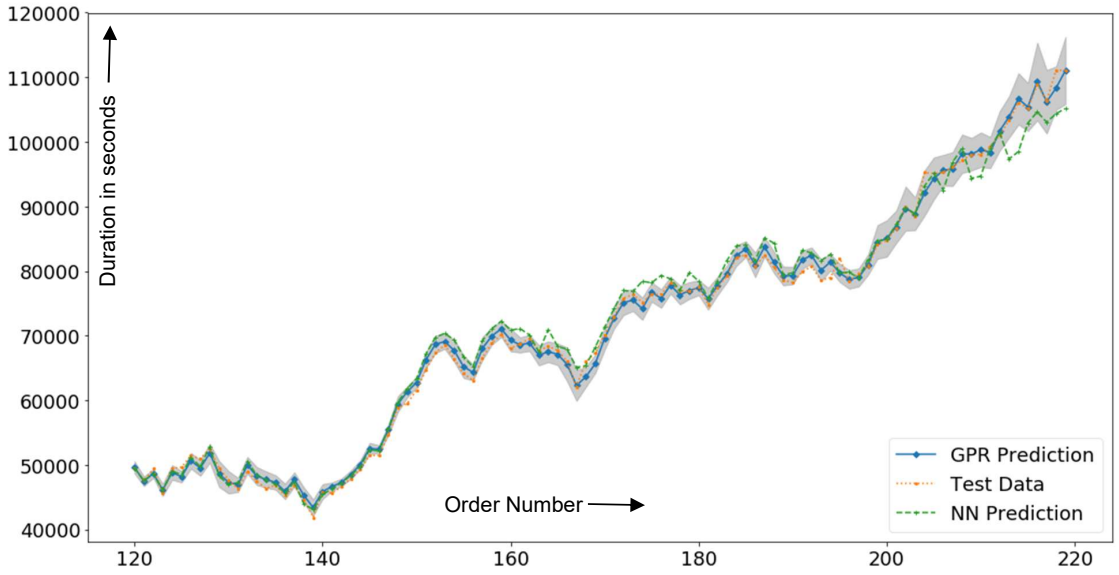


Figure 3: Order-completion duration prediction for test data at low load levels.

Generating data from simulations allows us to build and test these predictive models, and pay close attention to different system conditions, such as light load versus heavy load. Such comparisons are harder to make if we relied only on real factory data.

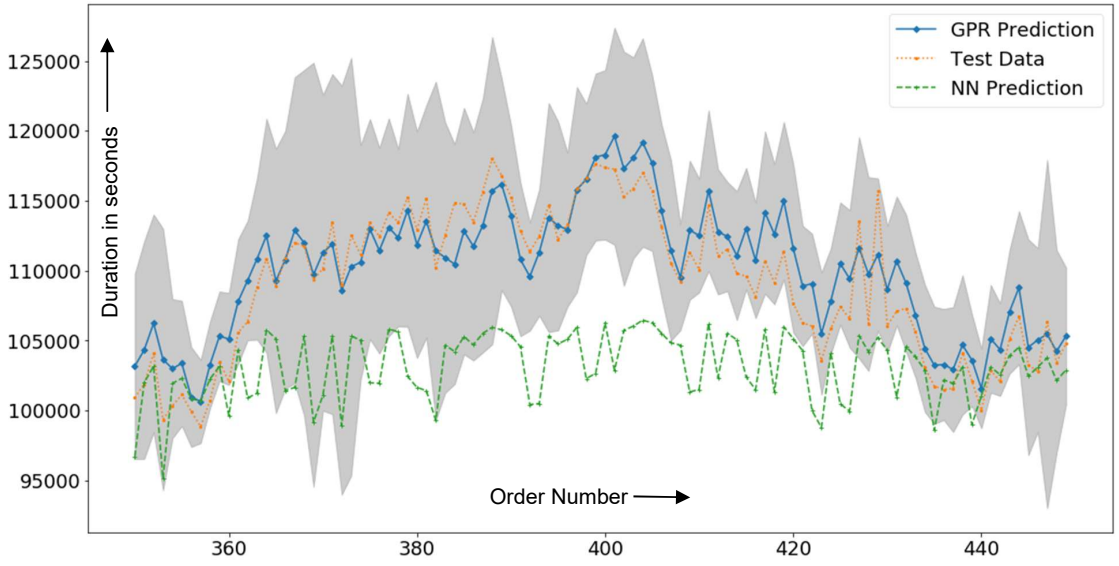


Figure 4: Order-completion duration prediction for test data at high load levels.

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

Our training data in the above example consisted of 800 training points from simulated data. As an experiment, we performed the training on a highly reduced training set of only 40 points (choosing every 20th in the original training set). Surprisingly, this reduced training set produced good predictions on the same data set, with only a slightly larger error. The GPR model had RMSE of 2938 seconds, while the NN model had an RMSE of 8670 seconds. The GPR model performed significantly better on the reduced training set. The error in the predicted duration is still quite small, compared to the average order-completion duration of 61426 seconds. Figure 5 shows the reduced training set. The important point to note is that the reduced training set still has good coverage over the range of duration values observed. The predictions for corresponding models are shown in Figure 6. In real factory situations, getting quality training data may often prove difficult. Simulated data such as this can be used to quickly evaluate the performance of different models, and make an appropriate choice of model to train on the limited factory data. The performance of the models has been summarized in Table 1.

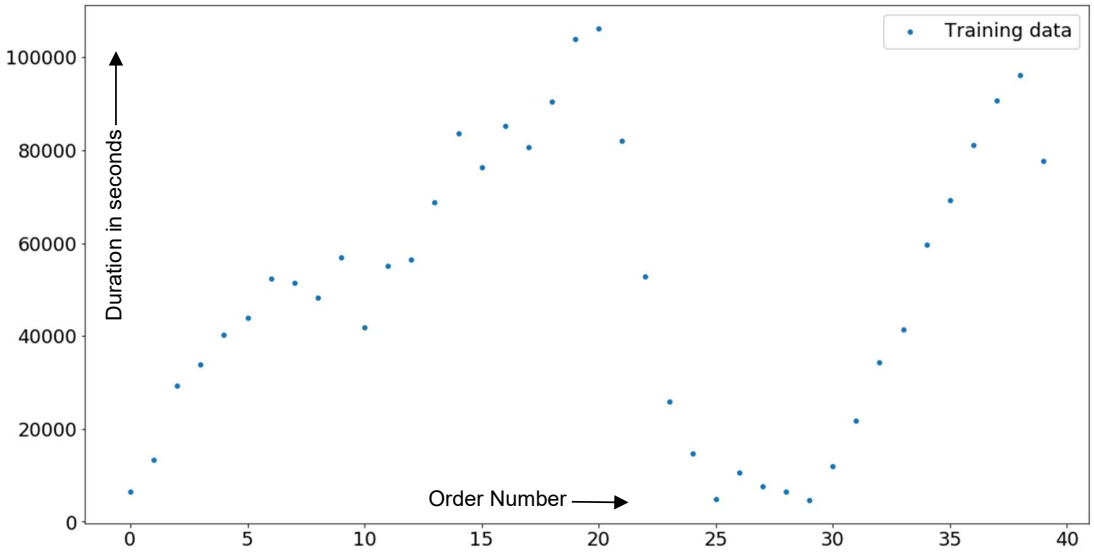


Figure 5: Reduced training data set.

The training time for GPR models increases as the data sets become large, and can become computationally expensive over large data sets. In our tests, the training times for the GPR model using the full training data and the limited training data were 75 milliseconds and 3 milliseconds, respectively. While the individual model training times may be improved by optimizing the code and improving system hardware, the relative times are indicative of the time savings when the model can be trained with less data. In our study, GPR proved to be a good model for training over smaller data sets, as long as the data sets had a good coverage over the expected operational range of the system parameters. In real factory scenarios, it may be difficult to quickly build a data set that has this kind of coverage. Our study using simulated data shows that GPR model performs very well with limited data. The simulations can be a guide to obtain good factory data over the required range of values, and train a good predictive model with limited data.

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

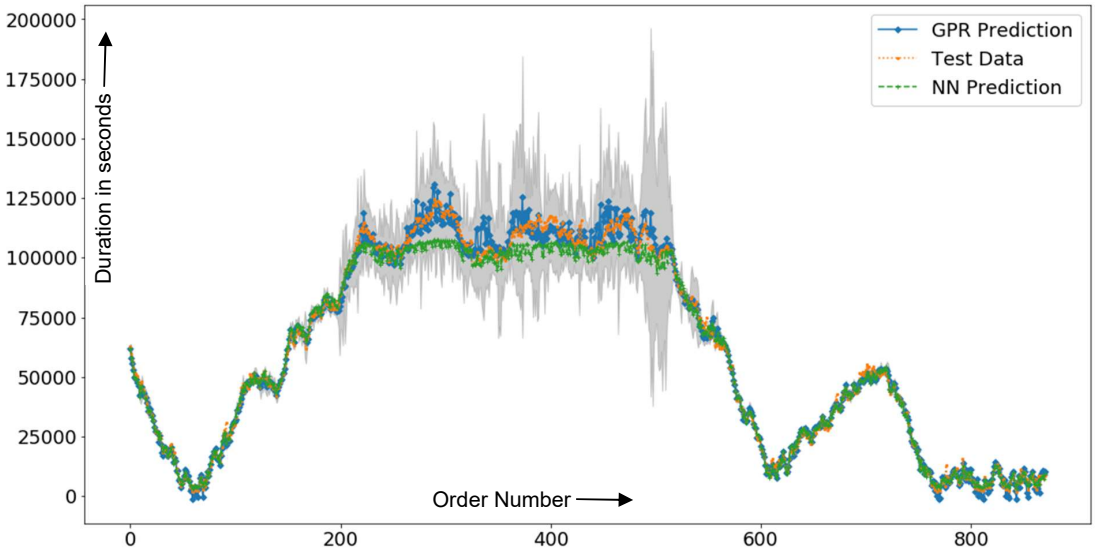


Figure 6: Predictions on reduced training data set.

Table 1: Performance of machine learning models.

Model	Data Set	RMSE	MAE
GPR	Full	1626	1241
NN	Full	4013	2414
GPR	Limited	2938	2008
NN	Limited	8670	6353

7 CONCLUSION

In this paper, we studied GPR as a machine learning model to predict throughput for a small job shop. We compared the GPR model with an NN model for the same purpose. A simulation of the job shop was used to generate data under varying load conditions, and the generated data was used to train and validate the machine learning models. Being able to generate synthetic data from simulation models allows us to build, test, and compare different machine learning models, which would be difficult to do if access to real factory data is limited. Our results showed that the GPR model performed better than the NN model, especially when the factory is operating under the high load condition. The GPR model also performed well when trained using limited data, while the NN model predictions were less accurate when trained with limited data. Testing these models using the simulation allows us to choose a machine learning model based on data availability and prediction accuracy. This allows manufacturers to build a data analytics and decision guidance system when real factory data is limited or not yet available. The GPR model also provides confidence bounds for its prediction which can help decision making.

Future directions under consideration include testing the DA and ML approaches for models of larger manufacturing systems, enhancing the virtual factory prototype for the purpose, generating a benchmark data set for comparing DA and ML approaches, and testing additional DA and ML approaches.

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

DISCLAIMER

No approval or endorsement of any commercial product by the National Institute of Standards and Technology (NIST) is intended or implied. Certain commercial software systems are identified in this paper to facilitate understanding. Such identification does not imply that these software systems are necessarily the best available for the purpose.

ACKNOWLEDGMENTS

The work of Sanjay Jain and Anantha Narayanan on this effort was supported by National Institute of Standards and Technology Cooperative Agreement Nos. 70NANB18H010 and 70NANB14H250, respectively.

REFERENCES

Ahmed, N. K., A. F. Atiya, N. E. Gayar, and H. El-Shishiny. 2010. "An Empirical Comparison of Machine Learning Models for Time Series Forecasting." *Econometric Reviews* 29(5-6):594-621.

Bhinge, R., N. Biswas, D. Dornfeld, J. Park, K. H. Law, M. Helu, and S. Rachuri. 2014. "An Intelligent Machine Monitoring System for Energy Prediction Using a Gaussian Process Regression." In *Proceedings of the 2014 IEEE International Conference on Big Data*, 978-986. Piscataway, NJ: IEEE.

Ding, D., C. Shen, Z. Pan, D. Cuiuri, H. Li, N. Larkin, and S. van Duin. "Towards an Automated Robotic Arc-Welding-Based Additive Manufacturing System from Cad to Finished Part." *Computer-Aided Design* 73(C):66-75.

Goebel, K., B. Saha, and A. Saxena. 2008. "A Comparison of Three Data-Driven Techniques for Prognostics." In *Failure Prevention for System Availability: Proceedings of 62nd Meeting of the Society for Machinery Failure Prevention Technology*, 119-131. Dayton, OH : Society for Machinery Failure Prevention Technology (MFPT).

Haykin, S. 2004. "NEURAL NETWORKS: A Comprehensive Foundation", Prentice Hall.

IIF. 2017. M3 Competition. International Institute of Forecasters. https://forecasters.org/resources/time-series-data/m3-competition/, accessed August 4th, 2018.

ISO. 2007. *ISO 10303-238:2007. Industrial Automation Systems And Integration.* Geneva, Switzerland: International Standards Organization.

Jain, S., G. Shao, and S.-J. Shin. 2017. "Manufacturing Data Analytics Using a Virtual Factory Representation". *International Journal of Production Research* 55(18):5450-5464.

Jain, S., D. Lechevalier, and A. Narayanan. 2017. "Towards Smart Manufacturing with Virtual Factory and Data Analytics." In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan et al., 3018-3029. Piscataway, NJ: IEEE.

Kamath, A., R. A. Vargas-Hernández, R. V. Krems, T. Carrington Jr, and S. Manzhos. 2018. "Neural Networks vs Gaussian Process Regression for Representing Potential Energy Surfaces: A Comparative Study of Fit Quality and Vibrational Spectrum Accuracy." *The Journal of Chemical Physics* 148(24): 241702.

Khorasani, A., and M. R. S. Yazdi. 2017. "Development of a Dynamic Surface Roughness Monitoring System Based on Artificial Neural Networks (ANN) in Milling Operation." *The International Journal of Advanced Manufacturing Technology* 93(1-4): 141-151.

Li, L., F. Liu, B. Chen, and C. B. Li. "Multi-Objective Optimization of Cutting Parameters in Sculptured Parts Machining Based on Neural Network." *Journal of Intelligent Manufacturing* 26(5): 891-898.

Liu, Y., and Z. Gao. 2015. "Real - Time Property Prediction for an Industrial Rubber - Mixing Process with Probabilistic Ensemble Gaussian Process Regression Models." *Journal of Applied Polymer Science* 132(6):41432.

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

MESA 2013. *Business To Manufacturing Markup Language Release Notes Version 6.0 – March 2013.* MESA International, Chandler AZ.

Miller, M. G., J. D. Summers, J. L. Mathieson, and G. M. Mocko. 2014. "Manufacturing Assembly Time Estimation Using Structural Complexity Metric Trained Artificial Neural Networks." *Journal of Computing and Information Science in Engineering* 14(1): 011005.

Rasmussen, C. E. 2004. "Gaussian Processes in Machine Learning." *Advanced Lectures on Machine Learning*. Springer, Berlin, Heidelberg, 63-71.

Rehman, M. A., D. Seth, and R. L. Shrivastava. "Impact of Green Manufacturing Practices on Organisational Performance in Indian Context: An Empirical Study." *Journal of Cleaner Production* 137 (2016): 427-448.

Scholz-Reiter, B., J. Heger, and T. Hildebrandt. 2010. "Gaussian Processes for Dispatching Rule Selection in Production Scheduling: Comparison of Learning Techniques." In *Proceedings of 2010 IEEE International Conference on Data Mining Workshops (ICDMW),* 631-638. Piscataway, NJ: IEEE.

SISO 2012. *SISO-STD-008-01-2012: Standard for Core Manufacturing Simulation Data – XML Representation.* Simulation Interoperability Standards Organization, Orlando, FL.

AUTHOR BIOGRAPHIES

SANJAY JAIN is an Associate Industry Professor in the Department of Decision Sciences, School of Business at the George Washington University. Before moving to academia, he accumulated over a dozen years of industrial R&D and consulting experience working at Accenture in Reston, VA, USA, Singapore Institute of Manufacturing Technology, Singapore and General Motors North American Operations Technical Center in Warren, MI, USA. His research interests are in application of modeling and simulation of complex scenarios including smart manufacturing systems and project management. His email address is jain@email.gwu.edu.

ANANTHA NARAYANAN is a Research Associate at the University of Maryland. He completed his PhD at Vanderbilt University, Nashville, TN, USA in 2008. His research interests are in data analytics and model based systems engineering. He is currently working as a guest associate in NIST's Systems Integration Division of the Engineering Laboratory. His email address is anantha@umd.edu.

Y. TINA LEE is a computer scientist in the Systems Integration Division of the Engineering Laboratory at NIST. Currently, she leads the Data Analytics for Smart Manufacturing Systems Project of NIST's Smart Manufacturing Systems Design and Analysis Program. She is the co-editor of the Simulation Interoperability Standards Organization (SISO) Standards of Core Manufacturing Simulation Data (CMSD). Her email address is yung-tsun.lee@nist.gov.

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun.
"COMPARISON OF DATA ANALYTICS APPROACHES USING SIMULATION."
Paper presented at 2018 Winter Simulation Conference, Gothenburg, Sweden. December 9, 2018 - December 12, 2018.

Condition-based Maintenance Policy Optimization Using Genetic Algorithms and Gaussian Markov Improvement Algorithm

Michael Hoffman[1], Eunhye Song[2], Michael Brundage[3], and Soundar Kumara[4]

[1,2,4] *The Pennsylvania State University, University Park, Pennsylvania, 16801, USA*
MichaelHoffman@psu.edu
eus358@psu.edu
u1o@engr.psu.edu

[3] *National Institute of Standards and Technology, Gaithersburg, Maryland, 20899, USA*
michael.brundage@nist.gov

ABSTRACT

Condition-based maintenance involves monitoring the degrading health of machines in a manufacturing system and scheduling maintenance to avoid costly unplanned failures. As compared with preventive maintenance, which maintains machines on a set schedule based on time or run time of a machine, condition-based maintenance attempts to minimize the number of times maintenance is performed on a machine while still attaining a prescribed level of availability. Condition-based methods save on maintenance costs and reduce unwanted downtime over its lifetime. Finding an analytically-optimal condition-based maintenance policy is difficult when the target system has non-uniform machines, stochastic maintenance time and capacity constraints on maintenance resources. In this work, we find an optimal condition-based maintenance policy for a serial manufacturing line using a genetic algorithm and the Gaussian Markov Improvement Algorithm, an optimization via simulation method for a stochastic problem with a discrete solution space. The effectiveness of these two algorithms will be compared. When a maintenance job (i.e., machine) is scheduled, it is placed in a queue that is serviced with either a first-in-first-out discipline or based on a priority. In the latter, we apply the concept of opportunistic window to identify a machine that has the largest potential to disrupt the production of the system and assign a high priority to the machine. A test case is presented to demonstrate this method and its improvement over traditional maintenance methods.

1. INTRODUCTION

The importance of maintenance in manufacturing is often overlooked as it is considered a non-value added activity in the manufacturing process. However, it is critical for supporting the availability and productivity of machines in the system. A maintenance policy defines how decisions are made regarding when and where to perform maintenance. In this work, we focus on the development of a condition-based maintenance policy for continuously monitored deteriorating machines. Since it is assumed that the health of each machine is known at all times, we can use this information to decide when maintenance should be performed.

In the work presented here, we consider non-uniform machines, a capacity for maintenance resources (a maximum number of concurrent maintenance jobs), and non-instant repair times. Much of the previous related work makes simplifying assumptions and does not consider all of these factors in combination. A capacity for maintenance resources and non-instant repairs results in frequent occurrences of conflicting maintenance jobs. This makes the development of a maintenance policy more difficult because, in addition to deciding when to repair each machine, we must decide how to reconcile maintenance scheduling conflicts. The objective when optimizing the maintenance policy is minimizing the cost of maintenance activities over some time horizon.

The rest of the paper is organized as follows: Section 2 describes some of the previous work related to maintenance policy optimization. Section 3 presents the system being considered. This includes the notation used, the machine deterioration model, the maintenance job queue, and the cost measurement model. The algorithms used to optimize the maintenance policy are described in Section 4. In Section 5 the results from an example are shown. Lastly, conclusions and future work are given in Section 6.

Hoffman, Michael; Song, Eunhye; Brundage, Michael; Kumara, Soundar.
"Condition-based maintenance policy optimization using genetic algorithms and Gaussian Markov improvement algorithm."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

2. BACKGROUND

Maintenance in manufacturing plays a pivotal role in ensuring efficiency in production. Maintenance includes all activities related to maintaining a specific level of availability of the system and components to perform at a certain level of quality and productivity (Al-Turki et al., 2014). One important aspect of maintenance is scheduling maintenance resources to ensure machine availability without sacrificing production throughput and quality. Scheduling maintenance resources requires determining when to send a technician, to what machine should the technician be sent, how often to schedule maintenance, and in what order, among other factors. For example, if two machines break down and a third machine is scheduled for preventive maintenance, which machine do you maintain first to avoid throughput loss and minimize cost?

Multiple maintenance scheduling policies exist that address this type of question. Jin et al. (2016) found that the majority of manufacturers still employ a mixture of reactive and preventive maintenance strategies. Reactive maintenance involves performing maintenance after an unanticipated failure of a piece of equipment. Preventive maintenance strategies involve performing maintenance after a set amount of time or after a piece of equipment is run for a certain number of cycles. These strategies are employed due to their low initial cost and low data requirements, however, they can lead to large productivity losses, equipment downtime, and can lead to over maintaining if the preventive maintenance is performed too frequently. To avoid over maintaining, condition-based maintenance (CBM) analyzes the current state of the machine and sets a predetermined level of health to maintain the machine (Jin et al., 2016). CBM requires large training data sets to build models for prediction, which requires considerable upfront cost and knowledge.

While the above strategies determine the frequency of maintenance, they do not address the priority rules for the order in which to perform the maintenance. This priority determination is especially challenging when maintenance resources are limited and several maintenance jobs are scheduled concurrently. The decision must be made as to which jobs should be carried out first. Opportunistic maintenance addresses this issue by analyzing trade-offs between production and maintenance to reduce production losses (Zhou, Xi, & Lee, 2009; Chang, Ni, Bandyopadhyay, Biller, & Xiao, 2007). This paper describes a CBM schedule with opportunistic priority rules minimizing both cost and production loss.

CBM optimization policies can be classified in several ways. Such classifications are based on maintenance policy parameters, system configuration, deterioration model, maintenance resource configuration, and optimization objectives (Khazraei & Deuse, 2011). This section provides examples of previous work in each classification scheme.

Two main categories of decision variables are considered when defining a CBM policy. One is the interval between inspections of components in the system. When continuous monitoring is not available, the health of components can only be known by performing an inspection that typically incurs some fixed cost. Upon inspection, the decision must be made as to whether or not maintenance should be performed. Such models are thoroughly described by Kallen & van Noortwijk (2006). The alternative, and the approach that is used here, is a continuously monitored system where the health of components is known at each discrete time step. In this case, the decision is related to at what health level should maintenance be scheduled.

The configuration of the system of interest will influence the optimal policy as well. Yang, Ma, & Zhao (2017) consider CBM of a single-unit system with multiple failure modes and determine the optimal policy given the state of the component. Maintenance policy optimization for multiple machines in series is studied by Bartholomew-Biggs, Zuo, & Li (2009). A series-parallel system is considered by Marseguerra, Zio, & Podofillini (2002), in which serial subsystems are comprised of identical machines in parallel. A policy is developed for a single machine type and then used for all machines within a subsystem.

A discrete-time Markov model is often used to represent the health of deteriorating machines. Each machine is assumed to be at a discrete level of health, known as the health index, at any point in time. The machine then transitions to another health index at the next time step with some probability. One variation of this model is the addition of random shocks, where a machine may transition into a complete failure state at any time, as examined by Yang et al. (2017). Some work has considered dependence among multiple machines, which can have a significant impact on the optimal maintenance policy. Rasmekomen & Parlikad (2016) develop a CBM policy for a system of components with stochastic dependence in which the degradation rate of a machine depends on that of others in the system.

While the minimization of cost is a typical objective in the design of maintenance policies, other competing objectives are also considered. In addition to minimizing cost, Marseguerra et al. (2002) aim to maximize the availability of the system. Lei, Liu, Ni, & Lee (2010) attempt also to maximize the throughput of the system by scheduling maintenance so that downtime does not hinder production. Many of these objectives can be combined into a single cost objective. For example, production that is lost due to downtime for maintenance can be assigned a cost that worsens the objective function value. By including such measures in the cost objective function, only one objective needs be considered.

Hoffman, Michael; Song, Eunhye; Brundage, Michael; Kumara, Soundar.
"Condition-based maintenance policy optimization using genetic algorithms and Gaussian Markov improvement algorithm."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

3. SYSTEM DESCRIPTION

In this section, we define the notation that is used throughout the remainder of the paper and the underlying assumptions of the system of interest. The system considered is a serial production line with M machines each with buffer size B and deterministic processing time t_m as depicted in Figure 1. Each machine will produce at its maximum rate while it is functional, so long as it is not starved or blocked. The first machine in the series is never starved and the last machine is never blocked.

3.1. Notation

- T - observed time horizon
- M - number of machines in series
- B - buffer capacity
- $b_m(t)$ - input buffer level of machine m at time t
- t_m - process time of a single part for machine m
- t_{m^*} - bottleneck process time (process time of bottleneck machine m^*)
- q_m - degradation rate of machine m
- $H_m(t)$ - health index of machine m at time t. $H_m(t) = 0$ indicates a machine is in perfect health, and the health index increases over time as the machine health deteriorates.
- h_m - health index threshold at which condition-based maintenance is scheduled for machine m
- $H_m(t) = h_{\max}$ - the health index at which a machine experiences total failure
- $\mathbf{P} = \{h_1, h_2, \cdots, h_M\}$ - CBM maintenance policy of the system
- r - maintenance capacity
- C_P - cost of a planned maintenance job
- C_U - cost of an unplanned maintenance job
- C_{LP} - cost per unit of lost production
- x_P - the number of planned maintenance jobs performed over the time horizon
- x_U - the number of unplanned maintenance jobs performed over the time horizon
- u - the actual number of units produced over the time horizon
- C_T - total policy cost

3.2. Deterioration Model

As described in Section 2, many condition-based maintenance processes assume that component deterioration can be modeled as a discrete Markov process. In this work, we assume that a machine m is in perfect working condition when its health index $H_m(t) = 0$ and that it degrades at each time step with a known probability. As the machine degrades, its health index increases until it is repaired or experiences a complete failure. The degradation rate of a machine can depend on many factors including age of the machine, stress on the machine, utilization, or the degradation state of other components in the system (Nicolai & Dekker, 2008). When a random failure occurs the machine stops functioning completely until it is repaired. When maintenance is completed on a component, whether preventive or in response to a complete failure, its health index is restored to zero and degradation resumes. The transition matrix $\mathbf{Q}_m$ of the degradation process of machine m is:

$$\mathbf{Q}_m = \begin{bmatrix} 1-q_m & q_m & & & & \\ & 1-q_m & q_m & & & \\ & & & \ddots & & \\ & & & & 1-q_m & q_m \\ & & & & & 1 \end{bmatrix} \quad (1)$$

Note that $\mathbf{Q}_m$ is upper bidiagonal.

Figure 2 shows the health index of a single machine over time. At time $t = 3$, the health index reaches the threshold for maintenance and is repaired at time $t = 4$. The machine is then restored to perfect health and degradation resumes. At time $t = 10$, the machine incurs a complete failure.

3.3. Maintenance Queue

Since we impose maintenance capacity on the system considered in this work, multiple maintenance jobs will simultaneously compete for limited maintenance resources. To handle these situations, we queue arriving maintenance jobs. When a maintenance job is placed in the queue (i.e. when the health index of a machine exceeds the threshold for CBM), it is serviced if there are sufficient maintenance resources available. Otherwise, it must wait for some time to be serviced. While waiting in the queue for maintenance, machines continue to degrade until their health index reaches $h_{\max}$.

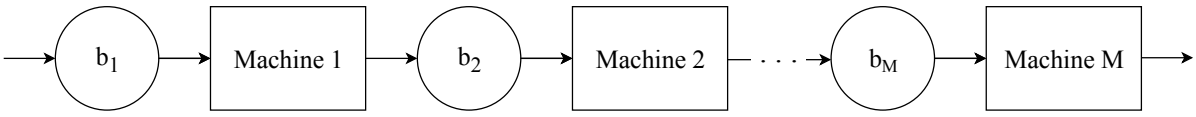


Figure 1. Series production system

Hoffman, Michael; Song, Eunhye; Brundage, Michael; Kumara, Soundar.
"Condition-based maintenance policy optimization using genetic algorithms and Gaussian Markov improvement algorithm."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

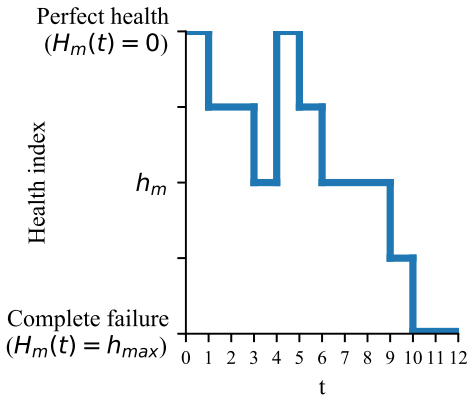


Figure 2. Degradation over time of a single machine, a repair is completed at time $t = 4$, and a complete failure occurs at time $t = 10$

We consider two queueing disciplines: first in, first out (FIFO) and priority queues. Under the FIFO rule, maintenance jobs are serviced in the order that they arrive in the queue by available maintenance resources. While this policy is simple to implement, it ignores the fact that high-risk machines with a greater potential to disrupt system throughput (e.g., the bottleneck machine) will be ignored if they are not in the front of the queue.

An alternative approach is to assign each maintenance job a priority measure and always service the job in the queue with the highest priority. To minimize lost production due to machine down time, maintenance jobs are assigned a priority that is related to the size of each machine's *maintenance opportunity window*. This concept is explained further in the following section.

3.3.1. Maintenance Opportunity Window

The maintenance opportunity window is the length of time a machine can stop production without hindering the overall system throughput. Throughput loss is avoided through the use of buffers in the system and by making sure the bottleneck machine is not blocked or starved. If a machine m in a serial line is upstream from the bottleneck machine m^* ($m < m^*$), the opportunity window for machine m is the time it takes for all buffers between m and m^* to become empty. At this point, the bottleneck machine is starved and throughput is hindered. If m is downstream from m^* ($m > m^*$), the opportunity window for m is the time for all buffers between m^* and m to become full. The bottleneck machine will then be blocked. This concept is described thoroughly by (Chang, Xiao, Biller, & Li, 2013) and summarized by Eq. (2):

$$W_m(t) = \begin{cases} t_{m^*} \sum_{k=m+1}^{m^*} b_k(t), & m < m^* \\ 0, & m = m^* \\ t_{m^*} \sum_{k=m^*+1}^{m} (B - b_k(t)), & m > m^*, \end{cases} \quad (2)$$

where $W_m(t)$ is the duration of the opportunity window for machine m at time t. This equation assumes that machine m is the only machine that is broken down over the duration of the opportunity window; however, work has shown how this equation also provides a rough estimate of opportunity window of each machine with simultaneous failures (Brundage, Chang, Li, Arinez, & Xiao, 2016). Future work will further refine the opportunity window equation for this purpose.

Machines with the smallest maintenance opportunity window will be assigned the highest priority. By minimizing the downtime of high-risk machines, we can reduce the throughput impact of performing maintenance. A comparison of the performance of the FIFO and priority queue policies is described in Section 5.

3.4. Cost Model

In general, minimizing the cost of the maintenance policy as given by Eq. (3) will be the primary objective. The cost of a policy over some time horizon consists of three components: planned maintenance activities ($C_P x_P$), unplanned maintenance activities ($C_U x_U$), and lost production due to downtime in the system (C_{LP}). The cost of a planned maintenance job is incurred when a machine is repaired before reaching the complete failure state. The total cost of planned maintenance is the product of the number of jobs that occur and the cost of planned maintenance activities. Similarly, the total cost of unplanned maintenance is the number of repairs completed on a machine in a failed state multiplied by the cost of the activity. The number of maintenance events over the observed time horizon is represented by a random variable.

Cost is also measured in terms of lost production due to machine downtime. Lost production is defined as the difference between the production requirement in units over the time horizon and the actual number of units produced by the system. The production requirement can be a fixed number of units, or a fraction of the "ideal production" that is obtained from a perfect system with no downtime events. The maintenance policy cost function is

$$C_T = C_P x_P + C_U x_U + C_{LP}\Big(\frac{T}{t_{m^*}} - u\Big). \quad (3)$$

Generally, planned maintenance is less costly than unplanned maintenance because the machine avoids a complete failure and downtime that results in system throughput disruption (Chitra, 2003). Unplanned maintenance occurs when a machine's health index reaches the total failure state, $h_{\max}$, and

Hoffman, Michael; Song, Eunhye; Brundage, Michael; Kumara, Soundar.
"Condition-based maintenance policy optimization using genetic algorithms and Gaussian Markov improvement algorithm."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

it is forced to stop production until repaired. Unplanned failures can occur when a machine that is waiting for planned maintenance continues to degrade to the point of total failure while waiting for maintenance resources to become available. Since we consider the duration of maintenance activities in the model, maintenance on a machine will disrupt the machine's production and possibly the overall throughput of the system. Lost production is defined as the difference between the production volume of the system if there was no degradation of machines and the actual production volume observed. Each unit of lost production incurs some cost that contributes to the cost of the maintenance policy.

4. METHODOLOGY

The goal of this work is to find an optimal CBM policy for a serial manufacturing system. As described in Section 2, a CBM policy is defined by the health index thresholds at which CBM is scheduled for each machine. Since there are M machines in the system, a solution, or policy, $\mathbf{P}$ can be value encoded by a set of M thresholds $\mathbf{P} = \{h_1, h_2, \cdots, h_M\}$. The minimum value of a threshold is 1, which would indicate maintenance is scheduled on a machine as soon its health index deteriorates by one unit. The maximum value is $h_{\max}$, the health index that indicates a machine has experienced a complete failure. A threshold at its maximum value is equivalent to a corrective maintenance policy.

For the problem presented here, the primary objective is to find the maintenance policy that minimizes expected cost per unit time. Due to the stochasticity and complex interactions that occur in the system under consideration, it is difficult to analytically determine the cost of the policy as a function of a set of CBM thresholds for each machine. Analytically determining cost often requires simplifying assumptions that reduce the accuracy of the cost measurement (Alrabghi & Tiwari, 2015). For this reason simulation will be used to evaluate the solutions by estimating the expected cost of a policy. We will compare the effectiveness of a genetic algorithm and the Gaussian Markov Improvement Algorithm in finding a solution to this problem In Section 4.3, details of the simulation model used in the experiments are given.

4.1. Genetic Algorithm

Genetic algorithms (GAs) are a metaheuristic method of problem solving that attempt to replicate evolutionary behavior found in nature. A population of solutions (or individuals) evolves over time by selecting the best members of the population to produce the succeeding generation. As in nature, a population benefits from diversity, so the algorithm begins with an initial set of random individuals. Starting with this initial population, offspring solutions are generated and added to the population. The best individuals from this group are then chosen to produce the next generation, and the process repeats until some termination criteria is met. There are four main considerations when using a genetic algorithm: the representation or encoding of the solutions within the context of the problem, the fitness function by which candidate solutions will be evaluated, the method of selecting individuals for reproduction, and the method of reproduction used (Deb et al., 2002). For this problem, a policy defined as $\mathbf{P} = \{h_1, h_2, \cdots, h_M\}$ represents an individual in the genetic algorithm.

The problem of optimizing a CBM policy is well-suited for GAs as this approach is robust and effective for large, complex manufacturing systems (Kobbacy, 2008).

4.2. Gaussian Markov Improvement Algorithm

Discrete optimization via simulation (DOvS) refers to finding an optimal solution of a problem with discrete decision variables whose objective function does not have a closed-form expression, but can be evaluated by stochastic simulation. Because of its flexibility, DOvS is a popular method for solving a complex stochastic problem. For the problems with small-to-medium feasible solution spaces where one can afford to simulate all feasible solutions, ranking and selection (R&S) has been successfully applied (Kim & Nelson, 2006); however, when the feasible solution space is large, it is practically impossible and inefficient to simulate all solutions. For the latter category of DOvS problems, several adaptive random search (ARS) algorithms have been developed. In general, an ARS algorithm initially simulates a small number of solutions and iteratively selects the next solution to simulate based on the simulation history. Since these initially sampled solutions are often used to estimate the necessary parameters of the algorithm, they are referred to as initial design points. A good ARS algorithm uses statistical inference based on simulated solutions to choose the next solution to simulate balancing exploration and exploitation.

The Gaussian Markov Improvement Algorithm (GMIA) finds the globally optimal solution of a DOvS problem with probability 1 when the simulation budget increases without a bound (Salemi, Song, Nelson, & Staum, 2018). GMIA is an ARS that draws statistical inference on the performance of feasible solutions by fitting a metamodel of the objective function at all feasible solutions based on the simulated solutions. The particular metamodel GMIA employs is a Gaussian Markov random field (GMRF), which models the unknown objective function values at the solution as Gaussian random variables with positive spatial correlations among solutions. From the simulation results of the initial design points, parameters of the GMRF model are estimated. Then, at each iteration, the distribution of the GMRF is updated conditional on the cumulative simulation results up to that iteration.

GMIA imposes the correlation such that nearby solutions have stronger positive correlation in the objective function

Hoffman, Michael; Song, Eunhye; Brundage, Michael; Kumara, Soundar.
"Condition-based maintenance policy optimization using genetic algorithms and Gaussian Markov improvement algorithm."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

Table 1. Test case 1 description.

Machine	1	2	3
t_m	3	4	5
q_m	0.02	0.06	0.01

values. This works for DOvS as solutions that are close in the feasible solution space often have similar objective function values. Therefore, even if a solution is not simulated yet, we can infer the objective function value at the solution based on simulated solutions and guide the search towards a more promising region of the feasible solution space. We defer the implementation details of GMIA in this paper; see (Salemi et al., 2018).

4.3. Simulation

Simulation is used to evaluate the quality of a maintenance policy solution. The system is simulated in its steady state for some period of time and then the cost of the defined maintenance policy is calculated using Eq. (3) described in Section 3.4. The system is considered to be in its steady state when the production rate of each machine is relatively constant over time. Once the steady state is achieved, the system is observed for the specified time horizon, T.

5. NUMERICAL RESULTS

A three-machine serial production line is used to demonstrate the methodology presented in the previous section. The system will be evaluated under both FIFO and priority queue disciplines. The machines in the system are described in Table 1. Table 2 describes additional parameters of the system.

GA and GMIA can be compared by evaluating the performance of each for a defined simulation budget. The simulation budget will be a maximum number of fitness function evaluations (NFE) that will occur. NFE for GMIA is given by

$$(2 \cdot \text{max iterations} + k) \cdot r, \tag{4}$$

where k is the initial number of design points and r is the number of simulation replications of each sampled solution. The NFE for GA depends on the population size, maximum number of generations, and the number of replications. NFE is given by

$$(2n + 1) \cdot \text{population size} \cdot \text{replications}, \tag{5}$$

where n is the number of generations. The parameters of both algorithms will be defined such that NFE is the same for both.

For the GMIA example shown here, the maximum number of iterations is 200, the number of initial design points is $k = 10$, and the number of simulation replications for each sampled solution is $r = 10$. This results in a total of 4100 fitness function evaluations at the termination of the GMIA. The parameters for the GA are a population size of 20, a maximum of 10 generations, and 10 simulation replications per solution. 4200 fitness function evaluations are used for GA.

For this system, the solution space is small enough that all solutions can be exhausted in order to find the global optimum. At 10,000 replications, the largest standard error observed was 3.23. The algorithms can also be compared to see if they converge to this solution.

5.1. FIFO Maintenance Queue

For this case, the overall best policy is $\mathbf{P} = \{8, 7, 8\}$ which was found to have a cost of 382.07 when simulated for 10,000 replications. Under this policy, machines 1, 2, and 3 are scheduled for repair when their health index reaches 8, 7 and 8, respectively. Figure 3 shows the convergence of each algorithm to the optimal solution for the system under a FIFO maintenance queueing discipline. The objective function value at each stage is the average of ten simulation replications. The cost shown on the vertical axis is the true cost of the best solution, as found by exhausting the solution space. When evaluating a solution, the algorithms are not likely to obtain an estimate that is equal to the true expected cost of the policy due to the high variance in the simulation. This results in the selection of "worse" solutions at some steps of each algorithm. Both algorithms are able to improve the solution over time, but on an average GMIA finds a policy with a lower cost.

In many cases, the true cost of a policy is different than that determined by the GA. As shown in Figure 4, the cost of the best solution at each generation as predicted by the GA (referred to as the observed cost) is much lower than the true cost of that solution. In fact, the observed cost of the best solution at termination is lower than the true minimum cost. This is

Table 2. Parameters for both test cases.

Parameter	Value
Buffer capacity (B)	2
Planned maintenance time to repair	Uniform$(5, 15)$
Planned maintenance cost (C_P)	50
Unplanned maintenance time to repair	Uniform(10, 20)
Unplanned maintenance cost (C_U)	300
Unit lost production cost (C_{LP})	10

Hoffman, Michael; Song, Eunhye; Brundage, Michael; Kumara, Soundar.
"Condition-based maintenance policy optimization using genetic algorithms and Gaussian Markov improvement algorithm."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

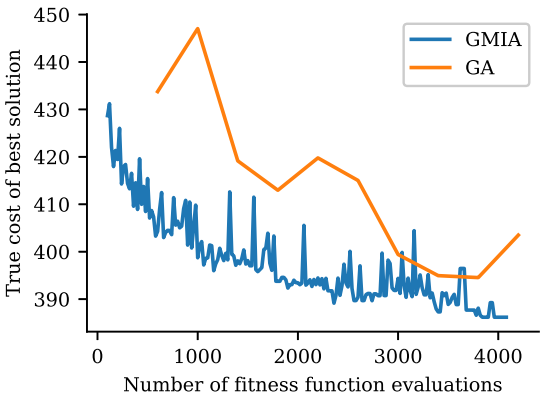


Figure 3. True cost of best solution versus simulation replications for a FIFO maintenance queue averaged over ten runs of each algorithm.

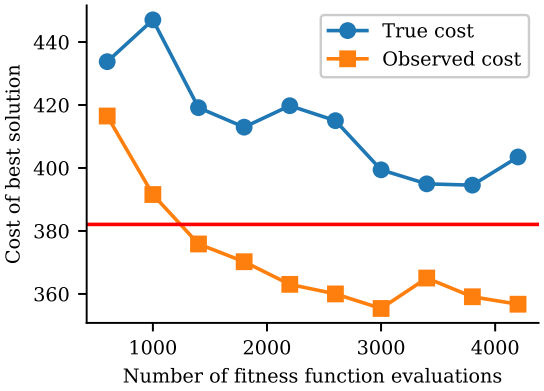


Figure 4. True and observed cost versus simulation replications for GA in a FIFO maintenance queue (horizontal line represents the true minimum cost)

due to the small number of simulation replications that are made when the GA evaluates a solution. The high degree of replication variability makes it difficult to accurately measure the fitness of a solution with only a few replications.

5.2. Priority Maintenance Queue

Similar results can be examined for the system under a priority queue discipline for maintenance jobs. The true minimum cost is obtained for the policy $\mathbf{P} = \{8, 7, 8\}$ which has an average objective function value of 383.21 after 10,000 simulation replications, so the cost is not improved by a priority queue. Both algorithms are again compared using a prescribed maximum number of fitness function evaluations. In Figure 5 the convergence of each algorithm is shown. Again it appears that on average the GMIA obtains a better solution for a given NFE.

Just as in the FIFO maintenance queue case, the GA tends to underestimate the cost of the best solution, as shown in Figure 6. This is again a result of the small number of replications that are used to evaluate the candidate solutions. There is a trade off between the accuracy of solution evaluations and the number of unique solutions evaluated. Conversely, another disadvantage of the GA is that favorable solutions may be overlooked due to the variability in their evaluation. Just as the fitness of some solutions is overestimated, it is likely that fitness is frequently underestimated as well. This could result in better solutions not being selected for reproduction, resulting in a non-optimal population of solutions.

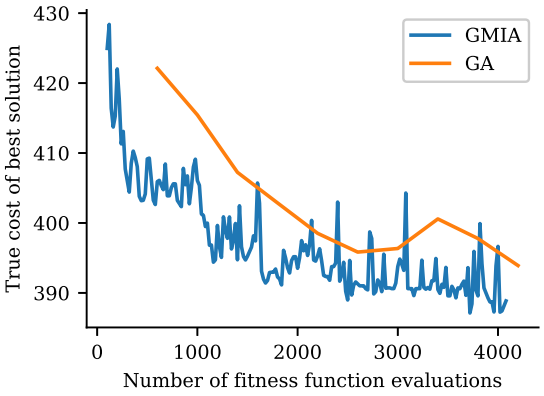


Figure 5. True cost of best solution versus simulation replications for a priority maintenance queue averaged over ten runs of each algorithm.

6. CONCLUSIONS

For the problem of optimizing a condition-based maintenance policy for a series manufacturing system, both GA and GMIA have shown to be effective search techniques. For a given

Hoffman, Michael; Song, Eunhye; Brundage, Michael; Kumara, Soundar.
"Condition-based maintenance policy optimization using genetic algorithms and Gaussian Markov improvement algorithm."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

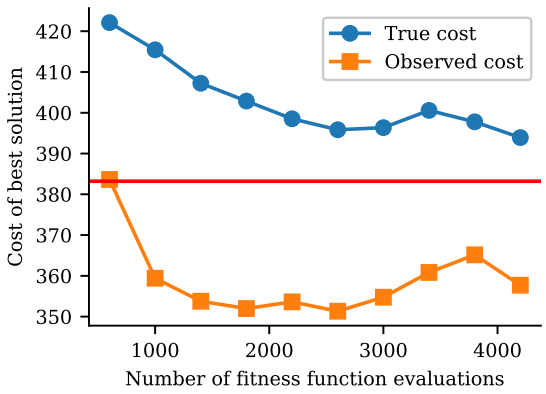


Figure 6. True and observed cost versus simulation replications for GA in a priority maintenance queue (horizontal line represents the true minimum cost)

simulation budget, GMIA is able to find a better optimal solution on average. This is an important consideration as the simulation of complex systems can be time-consuming and computationally expensive.

The maintenance of the example system presented here did not benefit from a priority maintenance job queueing discipline. This could be due to fact that failures of other machines are ignored when finding the opportunity window of a machine. Improving upon the opportunity window priority measure is among the next steps of this work. It may also be the case that there are not many instances of conflicting maintenance jobs, and so there is little need to decide the order in which jobs should be performed. A priority queue would likely be more effective for a system with a greater number of machines or machines with higher rates of degradation. In both cases, more maintenance jobs would occur over a given time horizon, thus increasing the occurrence of scheduling conflicts. This will be examined further in future work.

REFERENCES

Alrabghi, A., & Tiwari, A. (2015). State of the art in simulation-based optimisation for maintenance systems. *Computers Industrial Engineering*, *82*, 167.

Al-Turki, U. M., Ayar, T., Yilbas, B. S., & Sahin, A. Z. (2014). *Integrated maintenance planning in manufacturing systems*. Springer.

Bartholomew-Biggs, M., Zuo, M. J., & Li, X. (2009). Modelling and optimizing sequential imperfect preventive maintenance. *Reliability Engineering & System Safety*, *94*(1), 53–62.

Brundage, M. P., Chang, Q., Li, Y., Arinez, J., & Xiao, G. (2016). Implementing a real-time, energy-efficient control methodology to maximize manufacturing profits. *IEEE transactions on systems, man, and cybernetics: systems*, *46*(6), 855–866.

Chang, Q., Ni, J., Bandyopadhyay, P., Biller, S., & Xiao, G. (2007). Maintenance opportunity planning system. *Journal of Manufacturing Science and Engineering*, *129*(3), 661–668.

Chang, Q., Xiao, G., Biller, S., & Li, L. (2013). Energy saving opportunity analysis of automotive serial production systems (march 2012). *IEEE Transactions on Automation Science and Engineering*, *10*(2), 334-342.

Chitra, T. (2003). Life based maintenance policy for minimum cost. In (p. 470-474). IEEE.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, *6*(2), 182–197.

Jin, X., Siegel, D., Weiss, B. A., Gamel, E., Wang, W., Lee, J., & Ni, J. (2016). The present status and future growth of maintenance in US manufacturing: results from a pilot survey. *Manufacturing Rev.*, *3*, 10. doi: 10.1051/mfreview/2016005

Kallen, M. J., & van Noortwijk, J. M. (2006). Optimal periodic inspection of a deterioration process with sequential condition states. *International Journal of Pressure Vessels and Piping*, *83*(4), 249-255.

Khazraei, K., & Deuse, J. (2011). A strategic standpoint on maintenance taxonomy. *Journal of Facilities Management*, *9*(2), 96-113.

Kim, S.-H., & Nelson, B. L. (2006). Chapter 17 selecting the best system. In S. G. Henderson & B. L. Nelson (Eds.), *Simulation* (Vol. 13, p. 501 - 534). Elsevier.

Kobbacy, K. A. H. (2008). Artificial intelligence in maintenance. In *Complex system maintenance handbook* (pp. 209–231). London: Springer London.

Lei, Y., Liu, J., Ni, J., & Lee, J. (2010). Production line simulation using STPN for maintenance scheduling. *Journal of Intelligent Manufacturing*, *21*(2), 213-221.

Marseguerra, M., Zio, E., & Podofillini, L. (2002). Condition-based maintenance optimization by means of genetic algorithms and monte carlo simulation. *Reliability Engineering and System Safety*, *77*(2), 151-165.

Nicolai, R. P., & Dekker, R. (2008). Optimal maintenance of multi-component systems: A review. In (p. 263-286). London: Springer London.

Rasmekomen, N., & Parlikad, A. K. (2016). Condition-based maintenance of multi-component systems with degradation state-rate interactions. *Reliability Engineering System Safety*, *148*, 1-10.

Salemi, P., Song, E., Nelson, B. L., & Staum, J. (2018). Gaussian markov random fields for discrete optimization

Hoffman, Michael; Song, Eunhye; Brundage, Michael; Kumara, Soundar.
"Condition-based maintenance policy optimization using genetic algorithms and Gaussian Markov improvement algorithm."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

via simulation: Framework and algorithms. *Operations Research*.

Yang, L., Ma, X., & Zhao, Y. (2017). A condition-based maintenance model for a three-state system subject to degradation and environmental shocks. *Computers & Industrial Engineering*, *105*, 210-222.

Zhou, X., Xi, L., & Lee, J. (2009). Opportunistic preventive maintenance scheduling for a multi-unit series system based on dynamic programming. *International Journal of Production Economics*, *118*(2), 361–366.

BIOGRAPHIES

Michael Hoffman is a Ph.D. candidate in Industrial Engineering and Operations Research in the Department of Industrial and Manufacturing Engineering at the Penn State University. He is a Graduate Student Measurement Science and Engineering (GMSE) fellow at NIST. Previously, he was a Walker Graduate Assistant with the Applied Research Lab at Penn State. He received his B.S. in Industrial Engineering from Penn State. His research interests include intelligent manufacturing systems and big data in manufacturing.

Eunhye Song is Harold and Inge Marcus Early Career Assistant Professor in the Department of Industrial and Manufacturing Engineering at the Penn State University. She completed her B.S. and M.S. in Industrial and Systems Engineering from KAIST in Daejeon, South Korea and PhD in Industrial Engineering and Management Sciences from Northwestern University in Evanston, IL, USA. Her research interests include simulation design of experiments, simulation uncertainty and risk quantification, optimization via simulation under input model risk and large-scale discrete optimization via simulation.

Michael Brundage, Ph.D. is an Industrial Engineer in the Informational Modeling and Testing Group at the National Institute of Standards and Technology (NIST). Dr. Brundage's interests include Smart Manufacturing Diagnostics for Intelligent Maintenance, Sustainable Manufacturing Performance Measurement, Smart Manufacturing Capability Assessment, and Manufacturing Knowledge Visualization. His work contributes to guidelines for intelligent maintenance and he is part of a task group for creating an ASME Prognostics Health Management (PHM) standards committee. He also worked closely with ASTM International E60.13 in the development of a guideline for sustainable manufacturing performance indicators (ASTM E3096-17). He authored over 20 peer reviewed publications and has chaired multiple ASME MSEC Symposia and industry forums/workshops at NIST. Dr. Brundage is the recipient of the 2018 ASME Old Guard Early Career Award and was selected as one of SME's 2018 Class of 30 Under 30.

Soundar Kumara is the Allen, E., and Allen, M., Pearce Professor of Industrial Engineering at Penn State and has an affiliate appointment with the school of Information Sciences and Technology. His research interests are in smart manufacturing, large-scale networks, sensing and control, IIOT and Machine Learning in Manufacturing and Health Analytics. He is a Fellow of Institute of Industrial Engineers (IIE), International Academy of Production Engineering (CIRP), American Association for Advancement of Science (AAAS), and American Association of Mechanical Engineers (ASME). 54 Ph.D., and 64 MS students graduated under his tutelage. His Google citations is around 7700 and his Erdős number is 3.

Hoffman, Michael; Song, Eunhye; Brundage, Michael; Kumara, Soundar.
"Condition-based maintenance policy optimization using genetic algorithms and Gaussian Markov improvement algorithm."
Paper presented at 2018 Annual Conference of the Prognostics and Health Management Society, Philadelphia, PA, United States. September 24, 2018 - September 27, 2018.

Development of an Indoor Carbon Dioxide Metric

Andrew Persily

National Institute of Standards and Technology
100 Bureau Drive, MS8600
Gaithersburg, MD 20899 USA
**Corresponding author: andyp@nist.gov*

ABSTRACT

Indoor carbon dioxide (CO_2) concentrations have been used for decades to purportedly evaluate indoor air quality (IAQ) and ventilation. However, many applications of CO_2 as a metric have reflected a lack of understanding of the connection between indoor CO_2 levels, ventilation and IAQ. In many cases, an indoor concentration of 1800 mg/m^3 (1000 ppm_v) has been used as a metric of IAQ and ventilation without understanding its basis or significance. After many years of effort trying to dissuade practitioners as well as researchers from using this value, or some other concentration, as a metric of ventilation and IAQ, the author has developed an approach to determine a CO_2 level that can be used as a meaningful indicator of the outdoor ventilation rate per person. Rather than a single CO_2 concentration for all spaces and circumstances, this paper describes an approach to estimating a space-specific CO_2 concentration from several relevant factors. The concept is based on an estimate of the CO_2 concentration that would be expected in a specific space type given its intended or expected ventilation rate per person, the number of occupants and the rate at which they generate CO_2, and the occupancy schedule. A calculation method is described for estimating the CO_2 concentration for a given space and the timeframe for achieving that concentration, which provides a more meaningful metric than a single value for all spaces.

KEYWORDS

Building performance; carbon dioxide; indoor air quality; metrics; ventilation

1 INTRODUCTION

Indoor air quality (IAQ) is characterized by the chemical and physical constituents of air, plus other properties (e.g., thermal), that impact occupant health, comfort and productivity. The number of measurable airborne contaminants in most indoor environments is quite large, easily in the hundreds, and their impacts on building occupants is known for only a very small number. The large number of airborne contaminants, and their wide variation among buildings and over time, makes it extremely challenging to quantify IAQ conditions via a small number of parameters, let alone to distinguish between good and bad IAQ based on a single metric. There have been several efforts over the years to define IAQ metrics, but none have been shown to capture the occupant impacts of IAQ very well or have been accepted by the field (Hollick and Sangiovanni, 2000; Moschandreas et al., 2005; Jackson et al., 2011; Teichman et al., 2015).

Nevertheless, the indoor concentration of carbon dioxide (CO_2) has been widely promoted as a metric of IAQ and ventilation, in many cases without a clear understanding or explanation of what it is intended to characterize or a description of its application or limitations as a metric (Persily, 1997). At the simplest level, many practitioners use 1800 mg/m^3 (roughly 1000 ppm_v) as a metric, erroneously basing it on ASHRAE Standard 62.1 (ASHRAE, 2016a).

Persily, Andrew.
"Development of an Indoor Carbon Dioxide Metric."
Paper presented at 39th AIVC Conference: "Smart Ventilation for Buildings", Juan-les-Pins, France. September 18, 2018 - September 19, 2018.

Despite numerous statements to the contrary, that standard has not contained an indoor CO_2 limit for almost 30 years (Persily, 2015a). The CO_2 concentration limit was removed based on the confusion that it caused and the fact that it is not a good indicator of ventilation or IAQ. There have been many papers, presentations and workshops that have attempted to clarify the meaning of indoor CO_2 concentrations and even to advocate that they not be used as IAQ and ventilation metrics. However, it appears clear that calls to stop using indoor CO_2 to characterize IAQ and ventilation are not succeeding. Instead, efforts to educate designers, practitioners and others in the field need to continue, and this paper proposes an approach to using indoor CO_2 concentrations as a metric of ventilation rate per person based on a thorough consideration of the relevant parameters that determine indoor CO_2 levels.

2 BACKGROUND ON INDOOR CO_2 CONCENTRATIONS

Indoor CO_2 concentrations have been prominent in discussions of ventilation and IAQ since the 18th century, when Lavoisier suggested that CO_2 build-up rather than oxygen depletion was responsible for "bad air" indoors (Klauss et al., 1970). About one hundred years later, von Pettenkofer suggested that biological contaminants from human occupants were causing indoor air problems, not CO_2. Since that time, discussions of CO_2 in relation to IAQ and ventilation have evolved, focusing on the impacts of CO_2 concentrations on building occupants, how these concentrations relate to occupant perception of bioeffluents, the use of indoor CO_2 concentrations to estimate ventilation rates, and the use of CO_2 to control outdoor air ventilation rates (Persily, 2015b).

Indoor CO_2 concentrations are certainly relevant to the outdoor air ventilation rates per person specified in standards, guidelines and building regulations (CEN, 2007; ASHRAE, 2016a; ASHRAE, 2016b). These outdoor air requirements reflect more than 100 years of research, which first focused on the amount of ventilation needed to control odor associated with the byproducts of human metabolism, i.e., bioeffluents (Klauss et al., 1970). This research found that about 7.5 L/s to 9 L/s per person of ventilation air diluted body odor to levels judged to be acceptable by individuals entering the room from relatively clean air, i.e., unadapted visitors. Some of these experiments also included measurements of CO_2 concentrations, allowing examination of the relationship between CO_2 concentrations and body odor acceptability. The finding that about 8 L/s per person of ventilation controlled human body odor such that about 80 % of unadapted visitors found the odor to be acceptable was accompanied by the result that the same level of acceptability occurred at CO_2 concentrations about 1200 mg/m^3 above outdoors. For an outdoor CO_2 level of 600 mg/m^3, this concentration difference corresponds roughly to the commonly-cited indoor value of 1800 mg/m^3. (Note that outdoor levels have increased to 700 mg/m^3 or more since these odor acceptability studies were done (NOAA, 2018).) This body of research supports 1800 mg/m^3 of CO_2 as a reflection of body odor acceptability perceived by unadapted visitors to a building. Of course, there are many other important indoor air contaminants that are not associated with the number of occupants, and CO_2 concentration is not a good indicator of those contaminants.

ASHRAE Standard 62-1981 contained an indoor CO_2 limit of 4500 mg/m^3 for use when applying the performance approach to complying with the standard, i.e., the Indoor Air Quality Procedure. That limit was changed without written explanation to 1800 mg/m^3 in the 1989 version of the standard. That value was viewed by many a de facto standard without a sound understanding of its basis as an indicator of body odor acceptability to unadapted building occupants (Persily, 1997). This 1997 reference notes the existence of anecdotal discussions associating CO_2 concentrations in this range with occupant symptoms such as

Persily, Andrew.
"Development of an Indoor Carbon Dioxide Metric."
Paper presented at 39th AIVC Conference: "Smart Ventilation for Buildings", Juan-les-Pins, France. September 18, 2018 - September 19, 2018.

stuffiness and discomfort, also noting that peer-reviewed studies do not support these associations with the CO_2 itself. While several studies have shown associations of elevated CO_2 levels with symptoms, absenteeism and other effects (Apte et al., 2000; Shendell et al., 2004; Gaihre et al., 2014), these associations are likely due to lower ventilation rates elevating the concentrations of other more important indoor contaminants.

Indoor CO_2 concentrations are typically well below values of interest based on health concerns, though some recent work has shown evidence of impacts on human performance (Persily, 2015b). Two studies of individuals completing computer-based tests showed statistically significant decreases in decision-making performance at CO_2 concentrations as low as 1800 mg/m^3 (Satish et al., 2012; Allen et al., 2016). These experiments were carefully designed to expose the subjects to elevated CO_2 and not to other contaminants. However, other studies have not shown performance impacts at similar concentrations, therefore, it is premature to conclusively link CO_2 concentrations in this range with such occupant impacts (Zhang et al., 2016; Liu et al., 2017).

In summary, indoor CO_2 has not been shown to be a meaningful indicator of IAQ, and typical indoor levels do not have significant impacts on occupant health and comfort. Instead, this paper proposes using CO_2 as an indicator or metric of outdoor air ventilation rates per person. As discussed below, indoor CO_2 concentrations depend primarily on the rate at which the occupants generate CO_2 , the outdoor air ventilation rate of the space, the time since occupancy began, and the outdoor CO_2 concentration. Therefore, for indoor CO_2 to serve as a meaningful indicator of ventilation, all of these factors need to be considered.

2.1 Single-zone mass balance theory

The approach described in this paper, as well as many other discussions of indoor CO_2, employs a single-zone mass balance of CO_2 in the building or space of interest, which can be expressed as follows:

$$V\frac{dC}{dt} = Q\,(C_{out} - C) + G, \tag{1}$$

where V is the volume of the building or space being considered, C is the CO_2 concentration in the space in units of mg/m^3, C_{out} is the outdoor CO_2 concentration, t is time in hours, Q is the volumetric flow of air into the building (space) from outdoors and from the building (space) to the outdoors in m^3/h, and G is the CO_2 generation rate in the space in mg/h. Note that, in general, Q, C_{out} and G are functions of time, but they are assumed to be constant in this discussion. Also, air density differences between indoors and out are being ignored by using the same value of Q for the airflow into the space (building) and out. Finally, this single zone formulation ignores concentration differences between building zones and the CO_2 transport that occurs between zones. This last assumption is not always valid, and its appropriateness in any application of Equation 1 must be considered.

The solution to Equation 1 can be expressed as follows:

$$C(t) = C(0)e^{-\frac{Q}{V}t} + C_{ss}\left(1 - e^{-\frac{Q}{V}t}\right), \tag{2}$$

where $C(0)$ is the indoor concentration at $t = 0$ and C_{ss} is the steady-state indoor concentration. Note that the indoor concentration will only reach steady-state if conditions, specifically Q and G, are constant for a sufficiently long period of time, which can be many

Persily, Andrew.
"Development of an Indoor Carbon Dioxide Metric."
Paper presented at 39th AIVC Conference: "Smart Ventilation for Buildings", Juan-les-Pins, France. September 18, 2018 - September 19, 2018.

hours as discussed below. In particular, a constant value of *G* requires that the occupancy remain constant, and in many spaces occupancy will be too short or too variable for steady-state to be achieved. A convenient means of assessing whether steady-state is likely to be achieved is by considering the time constant of the system, which is equal to the inverse of *Q/V* in Equation 2, i.e., the inverse of the air change rate. One can consider that the system is essentially at steady-state after three time constants. For example, for a space with an air change rate of 1 h^{-1}, steady-state will exist after three hours. For a space with an air change rate of 0.5 h^{-1}, it will take six hours.

2.2 CO_2 generation from building occupants

The ventilation and IAQ fields have long used the following equation to estimate CO_2 generation rates from building occupants (ASHRAE, 2017):

$$V_{CO2} = \frac{0.00276 A_D M RQ}{(0.23RQ + 0.77)} \quad (3)$$

where V_{CO2} is the CO_2 generation rate per person (L/s); A_D is the DuBois surface area of the individual (m^2); *M* is the level of physical activity, sometimes referred to as the metabolic rate or met level (dimensionless); and *RQ* is the respiratory quotient (dimensionless). The respiratory quotient, *RQ*, is the ratio of the volumetric rate at which CO_2 is produced to the rate at which oxygen is consumed, and its value depends primarily on diet. Based on data on human nutrition in the U.S, specifically the ratios of fat, protein and carbohydrate intake, *RQ* equals about 0.85 (Persily and de Jonge, 2017).

More recently, an approach to estimating CO_2 generation rates from building occupants based on concepts from the fields of human metabolism and exercise physiology has been described (Persily and de Jonge, 2017). This approach uses the basal metabolic rate (*BMR*) of the individual(s) of interest, which is the energy needed to sustain the basic functions of human life, including the function of cells, the brain and the cardiac and respiratory systems, as well as the maintenance of body temperature. The *BMR* value of an individual is a function of their sex, age and body mass, which when multiplied by their level of physical activity or met level *M* yields their rate of energy expenditure. The rate of energy expenditure can then be related to oxygen consumption, and then CO_2 generation via the value of *RQ*. The noted reference provides equations to estimate *BMR* as well as data on met levels for different activities. Assuming RQ equals 0.85, the CO_2 generation rate of an individual can be estimated by the following equation:

$$V_{CO2} = BMR\ M\ 0.000484 \quad (4)$$

This updated approach for estimating CO_2 generation rates from individuals offers important advantages. First, Equation 3 is based on a 1981 reference that provides no explanation of its basis, while the new approach is derived using established principles of human metabolism and energy expenditure. Also, the new approach characterizes body size using mass rather than surface area, which in practice is estimated and not measured. Body mass is easily measured, and data on body mass distributions for various populations are readily available. The new approach also explicitly accounts for the sex and age of the individuals being considered, which is not the case with Equation 3.

3 CO_2-BASED VENTILATION METRIC

While a single CO_2 concentration metric that characterizes IAQ would be attractive, such a metric is not possible. As discussed earlier, there are many other indoor air contaminants with

Persily, Andrew.
"Development of an Indoor Carbon Dioxide Metric."
Paper presented at 39th AIVC Conference: "Smart Ventilation for Buildings", Juan-les-Pins, France. September 18, 2018 - September 19, 2018.

more significant health and comfort impacts than CO_2, and indoor CO_2 levels are rarely at concentrations of concern with respect to health effects. Instead, a CO_2 metric to evaluate outdoor air ventilation rates on a per person basis relative to a design value or a requirement in a standard is still of value, but it must be based on the space in question and its occupancy. The relevant space information includes the required outdoor air ventilation rate, its geometry (floor area and volume), and the number of occupants and their characteristics that impact the rate at which they generate CO_2 (sex, age, body mass and met level). This information can then be used to calculate the expected CO_2 concentration at a point in time, and that value can be related to a ventilation metric for a given space. However, performing such a calculation for each space is not realistic for many practitioners and applications. The approach taken in this paper is to perform these calculations using assumptions for the factors affecting CO_2 generation rates and ventilation rates. In order to explore these dependencies and how they relate to potential CO_2 metric values, indoor CO_2 concentrations were calculated for the space types listed in Table 1.

Table 1: Assumptions for CO_2 concentration calculations

		Outdoor air ventilation			
Space Type	**Occupant density** (#/100 m^2)	**L/s per person**	**h^{-1}**	**Occupants** (age, body mass in kg, met level)	**Average CO_2 generation per person (L/s)**
Classroom (5 to 8 y)	25	7.4	2.2	12 males (6 y, 23 kg, 2 met); 12 females (6 y, 23 kg, 2 met); 1 male (30 y, 85 kg, 3 met)	0.0043
Classroom (>9 y)	35	6.7	2.8	17 males (15 y, 68 kg, 1.7 met); 17 females (15 y, 61 kg, 1.7 met); 1 male (30 y, 85 kg, 2.5 met)	0.0059
Lecture classroom	65	4.3	3.3	32 males (20 y, 83 kg, 1.3 met); 32 females (20 y, 71 kg, 1.3 met); 1 male (30 y, 85 kg, 2.5 met)	0.0046
Restaurant dining room	70	5.1	4.3	33 males (30 y, 85 kg, 1.5 met); 33 females (30 y, 75 kg, 1.5 met); 2 males (30 y, 85 kg, 2 met); 2 females (30 y, 75 kg, 2 met)	0.0053
Conference meeting room	50	3.1	1.9	25 males (30 y, 85 kg, 1.3 met); 25 females (30 y, 75 kg, 1.3 met)	0.0044
Hotel/motel bedroom	10	5.5	0.7	5 male (30 y, 85 kg, 1 met); 5 female (30 y, 75 kg, 1 met)	0.0033
Office space	5	8.5	0.5	2.5 male (30 y, 85 kg, 1.4 met); 2.5 female (30 y, 75 kg, 1.4 met)	0.0047
Public assembly/Auditorium	150	2.7	4.9	75 males (30 y, 85 kg, 1.3 met); 75 females (30 y, 75 kg, 1.3 met)	0.0044
Public assembly/Lobby	150	2.7	4.9	75 males (30 y, 85 kg, 2 met); 75 females (30 y, 75 kg, 2 met)	0.0067
Retail/Sales	15	7.8	1.4	7.5 male (30 y, 85 kg, 2 met); 7.5 female (30 y, 75 kg, 2 met)	0.0067

Commercial/Institutional space types based on ASHRAE Standard 62.1-2016; outdoor air ventilation based on default occupancy density; ceiling height assumed to equal 3 m.

The space types considered in this analysis were selected from the longer list of commercial/institutional building space types in ASHRAE Standard 62.1 (ASHRAE, 2016a). Future analyses will consider residential buildings covered by Standard 62.2 and other standards (ASHRAE, 2016b), and perhaps other commercial/institutional space types. The second column of Table 1 is the occupant density, expressed as number of people per 100 m^2 of floor area (corresponding to the default values in Standard 62.1). The third and fourth

Persily, Andrew.
"Development of an Indoor Carbon Dioxide Metric."
Paper presented at 39th AIVC Conference: "Smart Ventilation for Buildings", Juan-les-Pins, France. September 18, 2018 - September 19, 2018.

columns are the outdoor air ventilation rate in L/s per person and h^{-1} based on Standard 62.1, with the conversion to h^{-1} using a ceiling height of 3 m. The fifth column contains information on the occupants (number, sex, age, body mass and met level) used to calculate their CO_2 generation rates, with the average per person generation rate in the last column. Most of the average CO_2 generation rates range from 0.004 L/s to 0.005 L/s. Higher values are seen for more active occupants, i.e., Public assembly/Lobby, Retail/Sales spaces and Classrooms (>9 y). A lower value of about 0.003 L/s is seen in the Hotel/motel bedroom spaces where the occupants are assumed to be sleeping, i.e., physical activity levels of 1 met.

For each space type the steady-state CO_2 concentration (relative to the outdoor level) and the time required to achieve steady-state were calculated using the assumptions listed in Table 1. These values are presented in the fourth and third columns in Table 2, along with the CO_2 concentration that would occur one hour after the space is fully occupied (in the fifth column). Also, a value of t_{metric} is listed for each space type in the second column of the table. This value is the length of time over which the particular space type may be expected to be fully occupied; the CO_2 concentration at that time is also listed in the table. These calculations assume all of the occupants enter the space at the same time, which is not necessarily the case in an actual building. The last three columns of the table contain the three CO_2 concentration values (steady-state, 1 h after full occupancy and t_{metric}) for a ventilation rate that is 25 % below the assumed value in Table 1. These reduced-ventilation cases are considered based on the desire for a CO_2-based ventilation metric to be able to capture ventilation deficiencies of this magnitude. The concentration calculations in this table employ the single-zone formulation in Equation 2 and therefore neglect any air and CO_2 transport from adjoining spaces. All of the input values used in these calculations can be revised in additional analyses. An online calculator is being developed to allow users to perform these calculations to examine the impact of different inputs.

Table 2: Calculated CO_2 concentrations

			CO_2 concentration above outdoors (mg/m^3)			**CO_2 for 25 % reduced ventilation rate (mg/m^3)**		
Space Type	**t_{metric} (h)**	**Time to steady-state (h)***	**Steady-state**	**1 h**	**t_{metric}**	**Steady-state**	**1 h**	**t_{metric}**
Classroom (5 to 8 y)	2	1.4	1060	940	1040	1410	1140	1360
Classroom (>9 y)	2	1.1	1580	1490	1580	2110	1860	2080
Lecture classroom	1	0.9	1940	1870	1870	2590	2370	2370
Restaurant	2	0.7	1871	1850	1870	2490	2390	2490
Conference room	1	1.6	2526	2130	2130	3370	2530	2530
Hotel/motel bedroom	6	4.5	1080	520	1060	1440	560	1370
Office space	2	5.9	985	390	630	1310	420	700
Auditorium	1	0.6	2900	2880	2880	3870	3770	3770
Lobby	1	0.6	4467	4430	4430	5960	5800	5800
Retail/Sales	2	2.1	1546	1170	1450	2060	1340	1810

* Time to achieve 95 % of steady-state CO_2 concentration, i.e., three time constants

The time to reach steady-state in Table 2 is linked to the air change rate in Table 1, i.e., it is three times the inverse of that rate. For most of the spaces, the time to steady-state is less than 1.5 h. In those cases, the three calculated CO_2 concentrations are generally within 100 mg/m^3, making the timing of a measurement for comparison to a metric less critical than in other spaces. For spaces with longer times required to achieve steady-state, the three calculated CO_2 concentrations cover a broader range. For these spaces, the concentration after 1 h of occupancy is more sensitive to the timing of the CO_2 measurement than the values at t_{metric} or at steady-state. It is worth noting that the concentrations at t_{metric} (and at steady-state and at 1 h for low time constant cases) tend to cluster around a discrete number of values: 600 mg/m^3,

Persily, Andrew.
"Development of an Indoor Carbon Dioxide Metric."
Paper presented at 39th AIVC Conference: "Smart Ventilation for Buildings", Juan-les-Pins, France. September 18, 2018 - September 19, 2018.

1000 mg/m^3, 1500 mg/m^3, 2000 mg/m^3, 3000 mg/m^3 and 4500 mg/m^3. Concentrations for other values of the inputs used in these calculations will likely be different, and the ability to identify characteristic concentration values will be reassessed after additional analyses. Of particular note is the Office space, which takes almost 6 h to reach steady-state due in large part to its low occupant density and low air change rate. As a result, the three concentrations values are all quite different. It's unlikely for a typical office space to be at full occupancy for 6 h given lunch schedules; therefore, the t_{metric} value of 2 h and the corresponding concentration of about 600 mg/m^3 are more relevant.

Consideration of the last three columns of Table 2 is useful for identifying a time at which the CO_2 concentration can be applied as a metric. As seen in this table, the CO_2 concentrations at t_{metric} generally exhibit a significant difference between the assumed ventilation rate and the 25 % ventilation deficiency. In cases where the time to reach steady-state is less than 1 h, the concentration at t_{metric} and at 1 h are essentially the same.

Based on the results in Table 2, and the desire to have a CO_2 metric that can capture ventilation deficiencies and be less sensitive to the timing of the concentration measurement, Table 3 summarizes potential CO_2 metric values for these spaces along with the corresponding measurement time. Given the transient nature of indoor CO_2 concentrations and the time to reach steady-state in many cases, it is not surprising that a potential CO_2 metric needs to be linked to a concentration measurement time. Therefore, reported CO_2 concentrations relative to these and other metrics need to include the time that has passed since the space reached full occupancy. Based on the analysis presented here, the time values are 1 h, 2 h and 6 h. A more complete analysis of other space types with different input values may yield other characteristic times. Future publications will present these additional analyses and an updated consideration of potential metric values.

Table 3: Potential CO_2 concentration metrics

Space Type	CO_2 concentration metric, above outdoors (mg/m^3)	Corresponding time (h after full occupancy)
Classroom (5 to 8 y)	1000	2
Classroom (>9 y)	1500	1
Lecture classroom	2000	1
Restaurant dining room	2000	1
Conference meeting room	2000	1
Hotel/motel bedroom	1000	6
Office space	600	2
Public assembly/Auditorium	3000	1
Public assembly/Lobby	4500	1
Retail/Sales	1500	2

The use of these concentration-time combinations as metrics of per person ventilation rates requires consideration of occupancy schedules. If the occupancy increases to the assumed full occupancy value over time, which is often the case, but one starts the calculation at the start of any occupancy, then the measured concentration at a given time will be less than the calculated value. Therefore, if the 1 h concentration value is used as a metric, the space could "pass" this criterion even though it would not do so over the long term. However, if the calculation doesn't start until full occupancy exists, then there would be some occupants in place before then, and the measured concentration would be "artificially" higher than it would be if occupancy started all at once. This situation would make the metric conservative, i.e., some spaces might "fail" even though they would pass if the space achieved full occupancy at a single instant in time. Note also that if the early occupants are different from the full occupants (in terms of CO_2 generation), it could be problematic.

Persily, Andrew.
"Development of an Indoor Carbon Dioxide Metric."
Paper presented at 39th AIVC Conference: "Smart Ventilation for Buildings", Juan-les-Pins, France. September 18, 2018 - September 19, 2018.

If the space is not at the occupancy level assumed in Table 1, which could easily be the case for retail or lobby spaces, one could estimate the fraction of the assumed occupancy and reduce the metric in Table 3 accordingly by multiplying by that fractional value. In fact, when applying this metric approach, the actual occupant density must be identified, and the concentration metric adjusted accordingly. It may not be practical to apply these metrics to spaces with particularly transient and short-term occupancies, such as retail and lobbies spaces, which speaks to the need to characterize the space occupancy and schedule as part of any such application.

Application of this CO_2 metric approach would require one to report, at a minimum, the following information: space type, occupant density, time at which full occupancy starts, time of CO_2 concentration measurement, and measured indoor and outdoor CO_2 concentrations. These measurements could then be compared with the values in Table 3, or a subsequent and more comprehensive version, as an indication of whether the ventilation rate per person complies with the value in Standard 62.1 or other ventilation requirement of interest. A more complete application of the approach could involve additional information, including: the ventilation rate per person target value (as an alternative to Standard 62.1), CO_2 concentration measurements at 15 min intervals starting at initial occupancy, and information on the ventilation strategy and system operation. As additional analyses are performed and the concept discussed with ventilation and IAQ practitioners and researchers, it is anticipated that the approach will become more well defined.

4 CONCLUSIONS

This paper presents an approach to using indoor CO_2 concentration measurements as a metric for ventilation rates per person, which accounts for the ventilation requirements and occupancies of specific space types. Calculations of steady-state CO_2 concentrations, as well as concentrations at other time intervals, are presented based on space-specific inputs of ventilation rate, space geometry and occupancy. These calculations are used to generate potential CO_2 concentration metrics for several space types in commercial/institutional buildings, along with measurement times after full occupancy that need to accompany CO_2 concentration measurements that are compared to these metrics. It is clear from these analyses that reported CO_2 concentrations for comparison to these, or any other metrics, need to be associated with a measurement time relative to the start of occupancy. Without information on time, such measurements cannot be interpreted.

Note that all of the input values used in these calculations can be revised to examine the impact of other values on the resulting CO_2 concentrations. An online calculator is being developed to allow users to perform these additional calculations. In addition, analyses are planned to study the concentrations in residential occupancies. These calculations will consider ventilation requirements from various international standards in single-family homes and multi-family units of different sizes.

5 ACKNOWLEDGEMENTS

The author expresses his appreciation to Steven J. Emmerich Kevin Y. Teichman and David A. Yashar for their helpful review comments.

6 REFERENCES

Persily, Andrew.
"Development of an Indoor Carbon Dioxide Metric."
Paper presented at 39th AIVC Conference: "Smart Ventilation for Buildings", Juan-les-Pins, France. September 18, 2018 - September 19, 2018.

Allen, JG, MacNaughton, P, Satish, U, Santanam, S, Vallarino, J and Spengler, JD. (2016). Associations of Cognitive Function Scores with Carbon Dioxide, Ventilation, and Volatile Organic Compound Exposures in Office Workers: A Controlled Exposure Study of Green and Conventional Office Environments. *Environ. Health Perspect.*, 124, 805-812.

Apte, MG, Fisk, WJ and Daisey, JM. (2000). Associations Between Indoor CO2 Concentrations and Sick Building Syndrome Symptoms in US Office Buildings: An Analysis of the 1994-1996 BASE Study Data. *Indoor Air*, 10 (4), 246-257.

ASHRAE. (2016a). *ANSI/ASHRAE Standard 62.1-2016 Ventilation for Acceptable Indoor Air Quality*, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA.

ASHRAE. (2016b). *ANSI/ASHRAE Standard 62.2-2016 Ventilation and Acceptable Indoor Air Quality in Low-Rise Residential Buildings*, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA.

ASHRAE. 2017. *Fundamentals Handbook,* Atlanta, GA, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.

CEN. (2007). *Ventilation for buildings - Energy performance of buildings - Guidelines for inspection of ventilation systems*, Brussels, European Committee for Standardization.

Gaihre, S, Semple, S, Miller, J, Fielding, S and Turner, S. (2014). Classroom Carbon Dioxide Concentration, School Attendance, and Educational Attainment. *Journal of School Health*, 84(9), 569-574.

Hollick, HH and Sangiovanni, JJ (2000) A Proposed Indoor Air Quality Metric for Estimation of the Combined Effects of Gaseous Contaminants on Human Health and Comfort, In: Nagda, N. L. (ed) *Air Quality and Comfort in Airliner Cabins, ASTM STP 1393*, West Conshohocken, PA, American Society for Testing and Materials, 76-98.

Jackson, MC, Penn, RL, Aldred, JR, Zeliger, HI, Cude, GE, Neace, LM, Kuhs, JF and Corsi, RL (2011) Comparison Of Metrics For Characterizing The Quality Of Indoor Air, *12th International Conference on Indoor Air Quality and Climate*, Austin, Texas.

Klauss, AK, Tull, RH, Roots, LM and Pfafflin, JR. (1970). History of the Changing Concepts in Ventilation Requirements. *ASHRAE Journal*, 12, 51-55.

Liu, W, Zhong, W and Wargocki, P. (2017). Performance, acute health symptoms and physiological responses during exposure to high air temperature and carbon dioxide concentration. *Building and Environment*, 114, 96-105.

Moschandreas, D, Yoon, S and Demirev, D. (2005). Validation of the Indoor Environmental Index and Its Ability to Assess In-Office Air Quality. *Indoor Air*, 15 (11), 874-877.

NOAA. (2018). *Trends in Atmospheric Carbon Dioxide Levels*, National Oceanic and Atmospheric Administration, Earth Systems Research Laboratory.

Persily, A. (2015a). Challenges in developing ventilation and indoor air quality standards: The story of ASHRAE Standard 62. *Building and Environment*, 91, 61-69.

Persily, AK. (1997). Evaluating Building IAQ and Ventilation with Indoor Carbon Dioxide. *ASHRAE Transactions*, 103 (2), 193-204.

Persily, AK (2015b) Indoor Carbon Dioxide Concentrations in Ventilation and Indoor Air Quality Standards, *36th AIVC Conference Effective Ventilation in High Performance Buildings*, Madrid, Spain, Air Infiltration and Ventilation Centre, 810-819.

Persily, AK and de Jonge, L. (2017). Carbon Dioxide Generation Rates of Building Occupants. *Indoor Air*, 27, 868-879.

Satish, U, Mendell, MJ, Shekhar, K, Hotchi, T, Sullivan, D, Streufert, S and Fisk, WJ. (2012). Is CO2 an indoor pollutant? Direct effects of low-to-moderate CO2 concentrations on human decision-making performance. *Environmental Health Perspectives*, 120, 1671-1677.

Persily, Andrew.
"Development of an Indoor Carbon Dioxide Metric."
Paper presented at 39th AIVC Conference: "Smart Ventilation for Buildings", Juan-les-Pins, France. September 18, 2018 - September 19, 2018.

Shendell, DG, Prill, R, Fisk, WJ, Apte, MG, Blake, D and Faulkner, D. (2004). Associations Between Classroom CO_2 Concentrations and Student Attendance in Washington and Idaho. *Indoor Air*, 14 (5), 333-341.

Teichman, K, Howard-Reed, C, Persily, A and Emmerich, S. (2015). *Characterizing Indoor Air Quality Performance Using a Graphical Approach*, National Institutte of Standards and Technology.

Zhang, X, Wargocki, P, Lian, Z and Thyegod, C. (2016). Effects of Exposure to Carbon Dioxide and Bioeffluents on Perceived Air Quality, Self-assessed Acute Health Symptoms and Cognitive Performance. *Indoor Air*, Accepted 23-Jan-2016.

Persily, Andrew.
"Development of an Indoor Carbon Dioxide Metric."
Paper presented at 39th AIVC Conference: "Smart Ventilation for Buildings", Juan-les-Pins, France. September 18, 2018 - September 19, 2018.

Solid Freeform Fabrication 2018: Proceedings of the 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference
Reviewed Paper

An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing

Shanshan Zhang[a], Brandon Lane[b], Justin Whiting[b], Kevin Chou[a]

[a]Industrial Engineering, University of Louisville

Louisville, KY, 40292

[b]Engineering Laboratory, National Institute of Standards and Technology

Gaithersburg, MD, 20899

Abstract

This study investigates the thermal conductivity of metallic powder in laser powder-bed fusion (LPBF) additive manufacturing. The intent is to utilize a methodology combining laser flash testing, finite element (FE) heat transfer modeling, and an inverse method to indirectly measure the thermal conductivity of nickel-based super alloy 625 (IN625) and titanium alloy (Ti64) powder used in LPBF processes. A hollow test specimen geometry was designed and built with LPBF enclosing the un-melted powder to mimic the powder bed conditions. The specimens were then flash heated in a laser flash system to measure their transient temperature response. Next, a developed FE model and a multi-point optimization algorithm were applied to inversely analyze the thermal transient, and extract the thermal diffusivity and conductivity of the powder enclosed in the specimens. The results indicate that the thermal conductivity of IN625 powder used in LPBF ranges from 0.65 W/(m·K) to 1.02 W/(m·K) at 100 °C and 500 °C, respectively. On the other hand, Ti64 powder has a lower thermal conductivity than IN625 powder, about 35 % to 40 % smaller. However, the thermal conductivity ratio of the powder to the respective solid counterpart is not much different between the two materials, about 4 % to 7 %, which is largely temperature independent.

Keywords: laser powder-bed fusion; laser flash; inverse method; thermal properties

1. Introduction

Laser powder-bed fusion (LPBF) additive manufacturing (AM) is a process that fabricates parts in a metal powder bed environment by powder-layer spreading and laser heating, melting, and solidifying layer-by-layer. The metal powder bed plays a significant role in the heat transfer phenomenon during LPBF, because heat dissipation to the ambient influences the rate of solidification of the molten metal, and therefore, the microstructure and mechanical properties of the build. In addition, accurate measurement of thermal properties of a powder bed in AM is essential for valid process modeling and predictions. While there are numerous publications regarding the thermal properties of common solid materials, little research has been reported regarding powder thermal properties in AM.

The transient hot-wire method was studied and used by Wei et al. to measure the thermal conductivities of commercial AM metal powders in a pressurized inert gas environment, and the authors reported that the heat dissipation of a powder bed was influenced by gas infiltration [1].

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

However, Gusarov et al. claimed that the thermal conductivity of gases at ambient pressure is substantially lower than that of metals and considered less important than other factors such as contacts between particles [2]. In addition, many researchers estimated the thermal conductivity of powder in powder-bed fusion AM using numerical approaches. Early work on evaluating the thermal conductivity of composite media can be derived from the Maxwell approach [3-6], which has been improved by the consideration of contacts between neighboring particles and gas in the pores. Some models have been developed to investigate the heat transport mechanism of a powder bed in AM and simulate the effective thermal conductivity. For example, Siu et al. and Slavin et al. both incorporated contact effects, such as the contact angle and the neck area between the neighboring particles for heat transfer in a powder bed, and conducted an analytical study to compute the powder thermal conductivity [7, 8]. Moreover, Singh et al. utilized an artificial neural network approach to predict the effective thermal conductivity of a porous system, which may contribute toward AM powder-bed studies [9]. Gong et al. incorporated powder thermal conductivity obtained from hot-disk based measurements and an analytical means into a 3D finite element (FE) thermal model to simulate the thermal field/history in powder-bed electron beam additive manufacturing [10].

Among different techniques for thermal diffusivity measurements, laser flash analysis, which was first developed by Parker et al. [11], is a widely used method for a wide variety of materials with a high precision. It uses the transient thermal response of a sample after a short heating pulse by a laser, then utilizes various heat transfer models to extract the thermal diffusivity from the measured response. For heterogeneous or anisotropic materials, more complex models may be required. Inverse heat transfer methods, in conjunction with laser flash technique, have been used to evaluate the thermal properties of thin coated films [12-16]. The solution of the analysis in these studies was based on the minimization of the least-squared errors between numerical model predictions and experimentally measured data, which was detailed in a publication from Ozisik [17]. Parker's theory of the flash method assumes one-dimensional heat transfer, without heat losses, and the homogeneity of the tested specimen. On the other hand, it is difficult to measure the thermal conductivity of metal powder, particularly with the size (approximately <50 μm) used in powder-bed fusion. With the inverse method approach, Cheng et al. developed and validated a combined experimental-numerical method to evaluate the powder thermal conductivity using laser flash testing and numerical heat transfer simulations [18]. The authors used additively fabricated hollow samples, with specially designed internal geometry, to enclose powder from LPBF. The internal geometry was designed to overcome an issue in which a gap occurred between the top shell and the internal powder, as reported in [19], which resulted in thermal insulations and complicated heat transport in the testing sample.

Continued from the previous work [18], the objective of this study is to analyze the temperature-dependent thermal conductivity of powder used in LPBF additive manufacturing. The test specimens of different designs, with enclosed powder, were laser-flash heated to obtain experimental thermal response at different temperatures, and the developed inverse methodology was employed to evaluate the temperature-dependent thermal conductivity of both nickel super alloy 625 (IN625) and titanium alloy (Ti64) powder materials.

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

2. Experimental details

2.1. Specimen design and fabrication

The test specimens were thin hollow disks built vertically by the LPBF process and to encapsulate powder during fabrication. In addition, internal cone features, either on the top or both the top and bottom sides of the hollow disks were included to ensure the contact between powder and the solid shells, preventing a large-area gap caused by powder settling [19]. As an example, Figure 1(a) is a photo of a fabricated two cones (0.5 mm height) sample. The radial cross-section of the sample model is shown in Figure 1(b). The overall dimensions of hollow disks are 25 mm in diameter and 3 mm in height with a shell of 0.5 mm thickness. The internal geometric feature had three different cone features: (1) both cones with a height of 0.5 mm (noted as 2Cone-0.5 throughout the paper), (2) both cones with a height of 0.25 mm (2Cone-0.25), and (3) one cone with a height of 0.5 mm on the top (1Cone-0.5). The dimensions of the cone-feature designs are shown in Figure 1(c) and Figure 1(d) shows the radial cross-section in the build orientation, i.e., a vertical build.

An EOS M270 system[1] was employed for sample fabrications. The powder materials used included both IN625 and Ti64. To achieve a full-density build, the process parameters suggested by the manufacturer were adopted for the solid shells. For IN625, the process parameter set was 195 W laser power and 800 mm/s scan speed [20] and the layer thickness was set as 40 µm. For Ti64, a laser power of 170 W and a scan speed of 1250 mm/s [21] were used, with a layer thickness of 30 µm. For both materials, the hatch spacing was 100 µm. No laser exposure was applied to the internal hollow section, as it was intended to encapsulate powder.

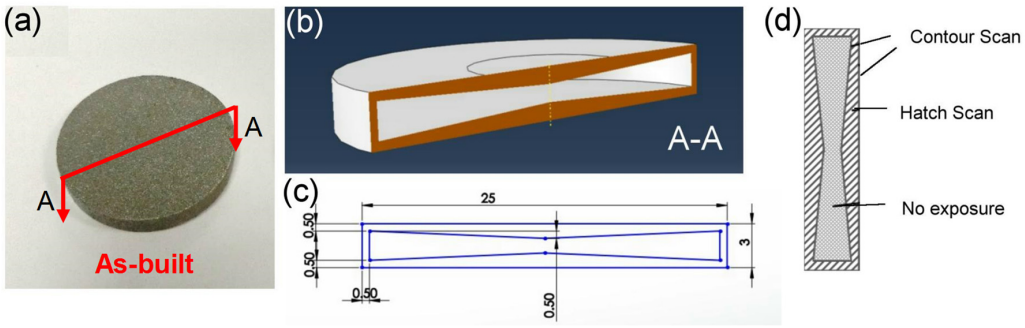


Figure 1. (a) An LPBF fabricated sample; (b) a geometric model of LPBF sample; (c) Dimensions of the 2Cone-0.5 powder-enclosed samples (unit: mm); (d) Build direction and scan conditions in LPBF.

2.2. Laser flash testing

Thermal diffusivity measurements of both solid and different encapsulated powder samples were carried out using a DLF-1200 from TA Instruments[1], shown in Figure 2(a). In this system, the test specimens are held in a furnace chamber, purged with either nitrogen or argon gas, which has environment temperature control that can be increased up to 1600 ºC. A laser pulse with a

[1] Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

variable energy up to 25 J was applied uniformly in a concentrically circular area with a diameter of about 22 mm at the bottom surface of the specimen. The laser power is adjusted and set automatically by the system to create an adequate thermal response and resulting signal from the pyrometer. The duration of laser irradiations was approximately 0.003 s. An infrared pyrometer collects thermal response from a 9.6 mm diameter circular region on the top surface of the test specimen and converts to digital signal output (Figure 2 (b)). To reduce laser reflection, the test specimen was coated with liquid graphite, and dried completely before loading onto a sample holder in the furnace chamber. During testing, the furnace heats to different programmed setpoint temperatures. Once steady state environment temperature is reached, the laser pulses, which increases the sample temperature only enough to enable a measurement from the pyrometer. The procedures and settings of specimen testing suggested by the manufacturer (TA Instruments) were followed.

The laser flash instrument generates a set of thermal radiation measurements collected over time via an infrared pyrometer. The experimentally acquired data is given as the voltage output and then transferred into a normalized response ranging from 0 to 1 which corresponds to the output at lowest and highest signal values. The response vs. time result is termed as a "thermogram." Since the diffusivity is related to the time response (e.g., rise time of the thermogram), knowledge of the absolute temperature rise due to the laser pulse is not necessary. The experimental results of IN625 and Ti64 are discussed in the following two sections.

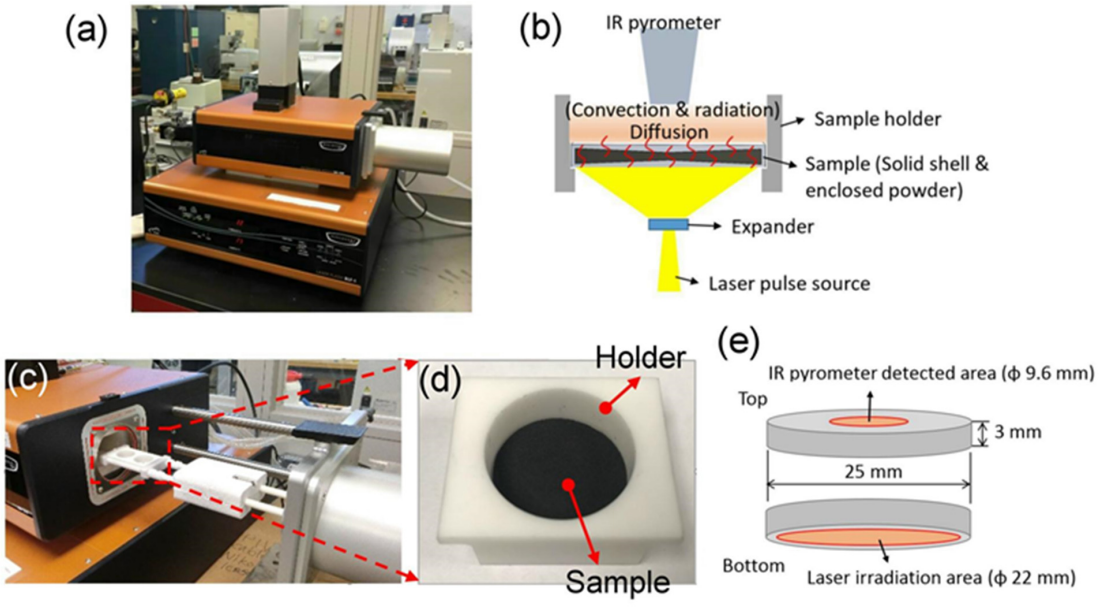


Figure 2. (a) DLF-1200 laser flash apparatus; (b) Schematic of laser flash method; (c) Specimen loading system; (d) Specimen holder; (e) Dimensions of IR detection and laser irradiation areas. Note that pyrometer spot size is not to scale as shown.

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

2.3. IN625 powder samples

Figure 3 shows the experimentally obtained thermograms of a solid sample as well as an example of the 2Cone-0.5 specimen with encapsulated powder. Compared with the solid sample, the heating rate of specimens with encapsulated powder is much slower; the maximum temperature is reached at between 10 s and 20 s vs. less than 3 s for the solid sample. It can also be noticed that as temperature increases, the heating period in the thermogram shifts to the left gradually due to increased thermal diffusivity of the IN625 material with the temperature.

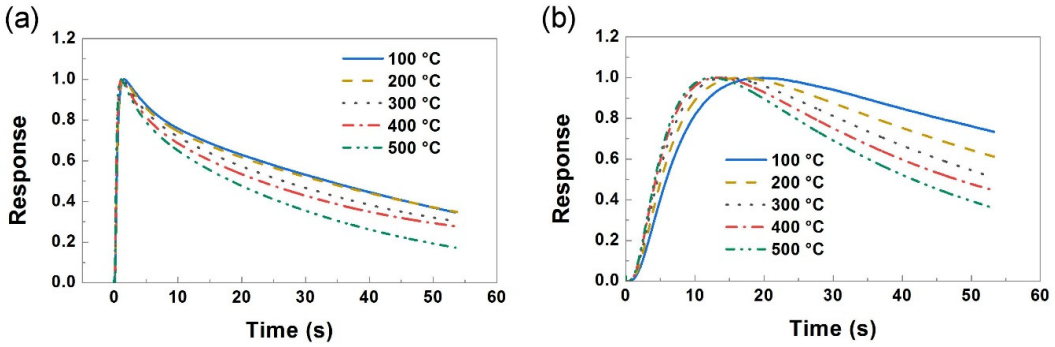


Figure 3. Experimental time-response thermograms of IN625 at various temperatures: (a) solid specimen measurement and (b) specimens with encapsulated powder (2Cone-0.5). Averages are taken from three separate measurements at each temperature.

Furthermore, at a given testing temperature, the thermograms of specimens with 2Cone-0.25 and 1Cone-0.5 features exhibit similar results in the heating period, and on the other hand, the 2Cone-0.5 specimen has a slightly higher heating rate than the specimens with the 2Cone-0.25 feature. An example of the comparison between the three cone features at 100 °C is shown in Figure 4.

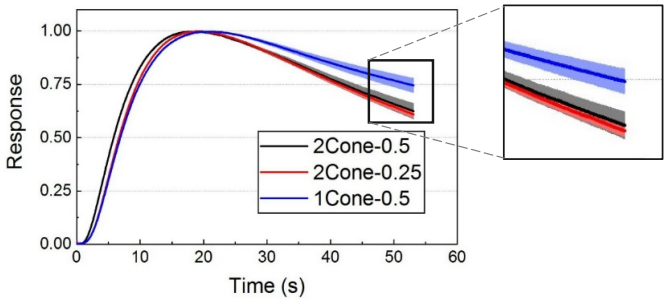


Figure 4. Comparison of laser flash thermograms of IN625 with three cone features at 100 °C. The light-color bars in the plot indicate the range of measurement results.

2.4. Ti64 powder samples

Same as the IN625 samples, thermograms from laser flash testing of Ti64 specimens show an increased thermal diffusivity as the testing temperature increases. Figure 5(a) shows the results of the 2Cone-0.5 specimen at various temperatures. Figure 5(b), on the other hand, compares

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

thermograms from laser flash testing of the IN625 and Ti64 specimens, both with encapsulated powder and with the 2Cone-0.5 feature. It is noted that the Ti64 specimen has slower heating compared to IN625 specimen, indicating a smaller diffusivity value because of the inherently lower thermal diffusivity of Ti64 alloy; in addition, the difference in thermograms between the two materials becomes smaller at higher temperatures.

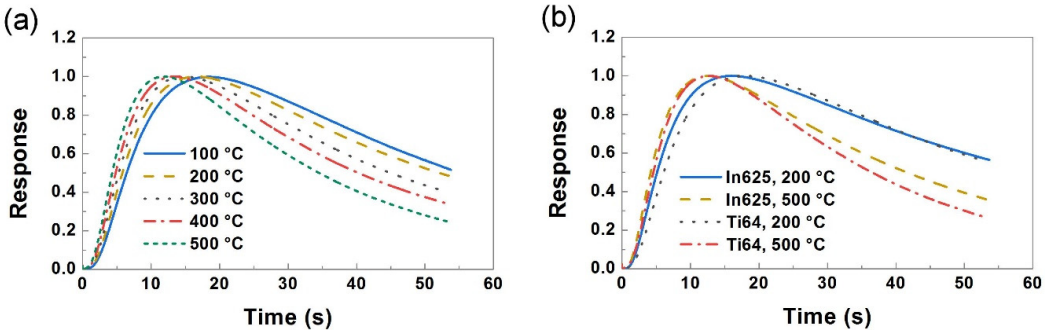


Figure 5. (a) Laser flash thermograms of Ti64 (Ti64) 2Cone-0.5 specimen; (b) Comparison of thermograms between IN625 and Ti64 specimens with encapsulated powder, both with 2Cone-0.5 feature.

3. Powder thermal conductivity evaluation

To analyze the thermal conductivity of the powder inside the LPBF-built specimens, the laser flash system was modeled and simulated by a finite element (FE) method using ABAQUS software. The specimen and its holder were modeled using the measured physical dimensions, with a mesh size of 0.5 mm and 0.7 mm, respectively. The laser heat source was simplified as a uniformly-distributed surface heat flux applied on the bottom side of the specimen. Convection and thermal radiation heat loss were included as the boundary conditions with the ambient temperature set as the testing temperature. The encapsulated powder was assumed to have the following unknown properties: density (ρ) and conductivity (k). Besides, two contact conductance values: (1) between the powder and the top solid shell (k_t), and (2) between the powder and the bottom solid shell (k_b), needed to be determined as well. Additionally, the specimen-holder contact conductance (k_p) at testing temperatures was obtained by analyzing the thermal response of the solid sample testing using the same laser flash system and the FE simulations, and then included in the laser flash simulation for the specimens with encapsulated powder.

Therefore, the problem is to accurately estimate the unknown LPBF powder thermal properties. A multivariate inverse method with a multi-point optimization algorithm was utilized to fit the simulation to the experimental results, and eventually to achieve the powder thermal conductivity in this study. The methodology of the inverse approach uses the Levenberg-Marquardt method, which has been used in a variety of inverse problems [17]. The complete approach, including FE simulations of laser flash testing and the inverse method, is detailed in a previous study [18].

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

3.1. IN625 powder study

In the thermal simulation of laser flash testing, temperature-dependent material properties of solid IN625 [22, 23] and alumina [24], which are applied for the solid capsule of the sample and the sample holder, respectively, are given in Figure 6. In addition, the density of alumina is assumed as 3800 kg/m^3 [25]. Moreover, the emissivity (unitless) for IN625 and alumina is 0.12 to 0.16 [26] and 0.7 [24], respectively. The convection coefficient was estimated to be 10 W/(m^2·K) [27]. The uncertainty of these assumed parameters is assumed to have an insignificant effect on the evaluation of the unknown parameters determined by the inverse method (e.g., powder thermal conductivity), although the sensitivity to parameter uncertainty is yet to be studied.

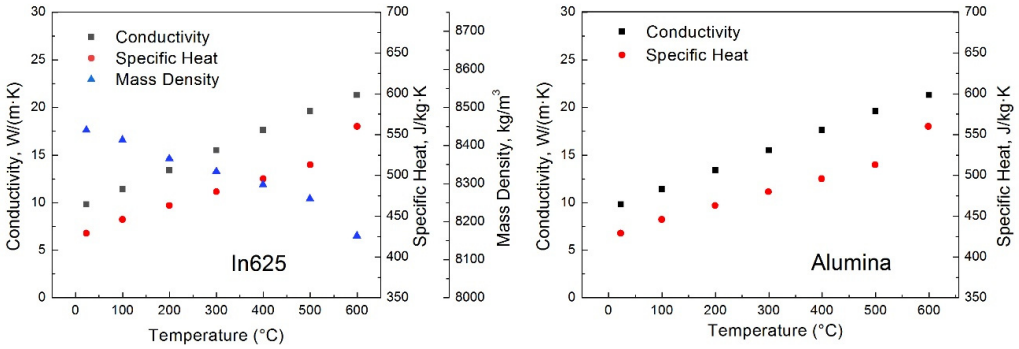


Figure 6. Material properties of solid IN625 sample and alumina sample holder.

3.1.1. Example of powder-enclosed sample analysis

Figure 7(a) shows the thermograms from three shots of laser flash testing of an IN625 specimen (2Cone-0.5) at 100 °C. To illustrate iterative results from the inverse method, Figure 7(b) shows the simulated thermogram from each iteration, with the third and fourth approaching the experimental curve. Table 1 below lists the simulations output as well as the overall error (S), calculated as the sum-squared error between the measured and simulated thermogram, calculated at each iteration. The initial values for the four unknowns were set as 10 % of the solid IN625 density and thermal conductivity at the testing temperature, and 100 W/(m^2·K) for the contact conductance. The initials were purposely set far away from the possible actual values to ensure no effects of initials to the final solution. By adjusting the damping factor in each iteration [18], an optimal set of the four unknown properties was selected for the next step. By calculating the S value (overall error), it can be determined if the simulation for the next iteration is necessary to proceed. In this case, the result from the 3rd iteration is considered the optimal solution, because the error increases at the 4th iteration.

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

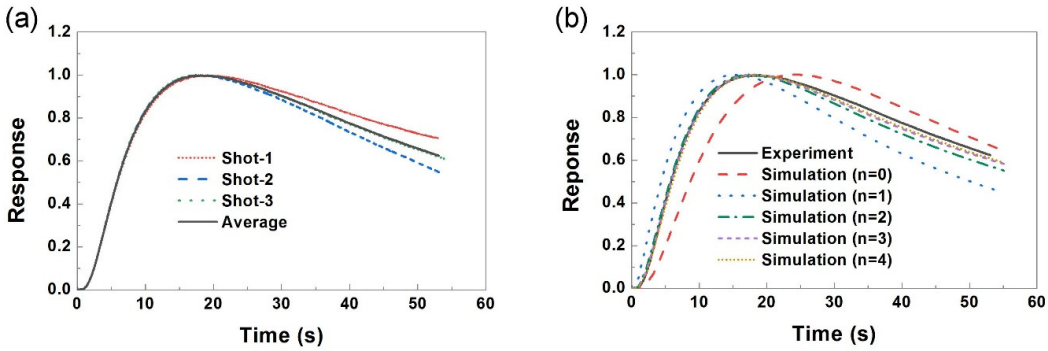


Figure 7. (a) Experimental results, including three shots and the average thermogram, and (b) Simulated thermograms from each iteration.

Table 1. Results from inverse method for IN625 2Cone-0.5 specimen at 100 °C.

n	Damping factor u	k, W/(m·K)	k_t, W/(m^2·K)	k_b, W/(m^2·K)	ρ, kg/m^3	S
0		0.1	100	100	841	0.512417
1	-2	0.4314	609.90	566.95	902.63	0.275936
2	-0.02996	0.7347	372.60	738.03	3682.66	0.015569
3*	0.05	0.7955	351.77	926.54	4775.56	0.003393
4	-9.2	0.8000	351.62	934.50	4740.68	0.003789

* Optimal solution

3.1.2. Temperature-dependent thermal conductivity

The laser flash results of the IN625 specimens with encapsulated powder, and three different cone features, at various temperatures were analyzed to inversely calculate the temperature-dependent powder conductivity. The results are summarized in Figure 8. The powder conductivity obtained ranges from 0.65 W/(m·K) to 1.02 W/(m·K), and generally, the powder conductivity is nearly linear to the temperature. However, the results extracted from the samples of three different cone features are slightly different. The models of the 2Cone-0.25 and 1Cone-0.5 give a similar powder thermal conductivity, while the powder conductivity analyzed from the 2Cone-0.5 model is about 0.1 W/(m·K) to 0.2 W/(m·K) higher than that from the other two models for all testing temperatures.

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

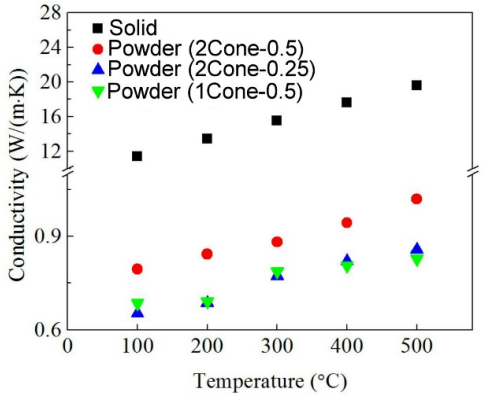


Figure 8. Thermal conductivity of IN625 powder.

3.2. Ti64 powder study

The FE model for Ti64 specimens with encapsulated powder was established also based on the actual geometry of fabricated specimens and Ti64 material properties. Figure 9 shows the material properties of solid Ti64 [28] that were incorporated in FE modeling. The same simulation approach and the inverse method used in the IN625 powder study were employed to the Ti64 2Cone-0.5 specimens with encapsulated powder.

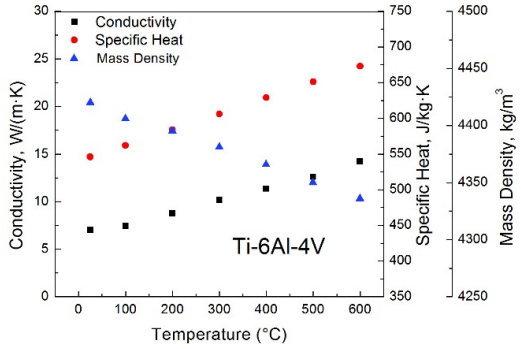


Figure 9. Thermal properties of solid Ti64.

The analyzed thermal conductivity values of Ti64 powder at various temperatures (100 °C to 500 °C) are plotted in Figure 10(a). It can be observed that the simulated Ti64 powder thermal conductivity linearly increases with temperatures and ranges from 0.39 W/(m·K) to 0.65 W/(m·K), for 100 °C and 500 °C, respectively. Moreover, when plotting in a normalized way (i.e., in reference to solid), it is noted that the Ti64 powder conductivity is approximately only 4 % to 5 % of the solid Ti64 conductivity at all testing temperatures, Figure 10(b). This finding is similar to the results of IN625 powder, which shows a slightly higher ratio, 4 % to 7 %, and different cone configurations result in a minor difference.

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

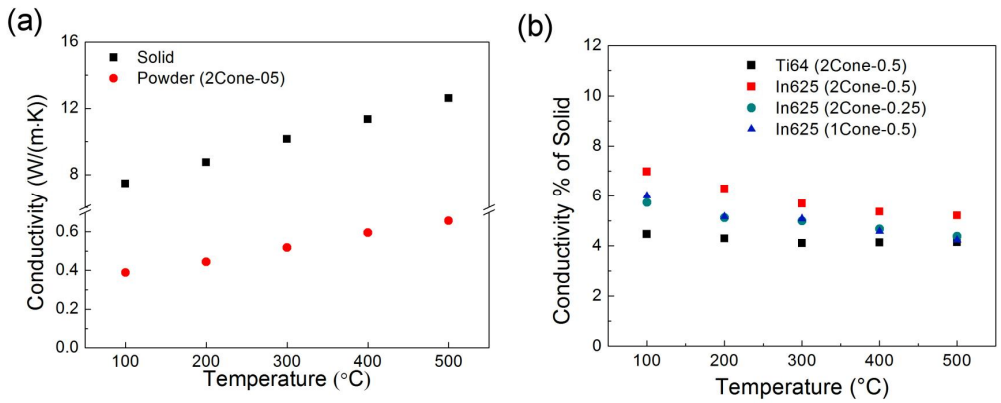


Figure 10. (a) Thermal conductivity of Ti64 powder at different temperatures, and (b) ratio of powder to solid thermal conductivities for IN625 and Ti64 (Ti64).

4. Conclusions

The LPBF specimens with encapsulated powder were designed and fabricated, to imitate powder-bed conditions in LPBF, by an EOS M270 system using two different powder materials: IN625 and Ti64. Different internal cone features were incorporated in the specimens to ensure contact between the powder and the top solid shell. To evaluate the powder thermal conductivity, laser flash experiments and a numerical approach using FE thermal simulations and an inverse method were conducted to analyze the powder thermal conductivity.

Based on the results obtained so far, it can be concluded that (1) the thermal conductivity of powder from LPBF is much lower than the solid conductivity, e.g., 0.65 W/(m·K) to 1.02 W/(m·K) for IN625, and 0.39 W/(m·K) to 0.65 W/(m·K) for Ti64, within the range of measured temperatures of 100 °C to 500 °C, (2) there is a linear correlation between the powder thermal conductivity value and the temperature, and (3) on the other hand, the powder thermal conductivity of both materials is approximately only 4 % to 7 % of their solid thermal conductivity, with Ti64 at a lower ratio.

5. Acknowledgements

This study is supported by National Institute of Standards and Technology (NIST) (CA No. 70NANB16H029). Discussions with and suggestions from Dr. Alkan Donmez are highly appreciated. The authors also acknowledge the technical support from Rapid Prototyping Center and the computational resource from Cardinal Research Cluster at University of Louisville.

6. References

1. Wei, L.C., L.E. Ehrlich, M.J. Powell-Palm, C. Montgomery, J. Beuth, and J.A. Malen, *Thermal conductivity of metal powders for powder bed additive manufacturing.* Additive Manufacturing, 2018. **21**: p. 201-208.
2. Gusarov, A., T. Laoui, L. Froyen, and V. Titov, *Contact thermal conductivity of a powder bed in selective laser sintering.* International Journal of Heat and Mass Transfer, 2003. **46**(6): p. 1103-1109.

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

3. Maxwell, J.C., *A treatise on electricity and magnetism*. Vol. 1. 1881: Clarendon press.
4. Nan, C.-W., R. Birringer, D.R. Clarke, and H. Gleiter, *Effective thermal conductivity of particulate composites with interfacial thermal resistance.* Journal of Applied Physics, 1997. **81**(10): p. 6692-6699.
5. Mercier, S., A. Molinari, and M. El Mouden, *Thermal conductivity of composite material with coated inclusions: applications to tetragonal array of spheroids.* Journal of Applied Physics, 2000. **87**(7): p. 3511-3519.
6. Gu, G. and Z. Liu, *Effects of contact resistance on thermal conductivity of composite media with a periodic structure.* Journal of Physics D: Applied Physics, 1992. **25**(2): p. 249.
7. Siu, W. and S.-K. Lee, *Effective conductivity computation of a packed bed using constriction resistance and contact angle effects.* International journal of heat and mass transfer, 2000. **43**(21): p. 3917-3924.
8. Slavin, A.J., F.A. Londry, and J. Harrison, *A new model for the effective thermal conductivity of packed beds of solid spheroids: alumina in helium between 100 and 500 C.* International Journal of Heat and Mass Transfer, 2000. **43**(12): p. 2059-2073.
9. Singh, R., R. Bhoopal, and S. Kumar, *Prediction of effective thermal conductivity of moist porous materials using artificial neural network approach.* Building and Environment, 2011. **46**(12): p. 2603-2608.
10. Gong, X., B. Cheng, S. Price, and K. Chou. *Powder-bed electron-beam-melting additive manufacturing: powder characterization, process simulation and metrology*. in *Proceedings of the ASME District F Early Career Technical Conference*. 2013.
11. Parker, W., R. Jenkins, C. Butler, and G. Abbott, *Flash method of determining thermal diffusivity, heat capacity, and thermal conductivity.* Journal of applied physics, 1961. **32**(9): p. 1679-1684.
12. Stryczniewicz, W. and A.J. Panas, *Numerical data processing from a laser flash experiment on thin graphite layer.* Computer Assisted Methods in Engineering and Science, 2017. **22**(3): p. 279-287.
13. Wright, L., X.-S. Yang, C. Matthews, L. Chapman, and S. Roberts, *Parameter estimation from laser flash experiment data*, in *Computational Optimization and Applications in Engineering and Industry*. 2011, Springer. p. 205-220.
14. McMasters, R.L., R.B. Dinwiddie, and A. Haji-Sheikh, *Estimating the thermal conductivity of a film on a known substrate.* Journal of Thermophysics and Heat Transfer, 2007. **21**(4): p. 681-687.
15. Kim, S.-K. and Y.-J. Kim, *Determination of apparent thickness of graphite coating in flash method.* Thermochimica Acta, 2008. **468**(1-2): p. 6-9.
16. Cernuschi, F., L. Lorenzoni, P. Bianchi, and A. Figari, *The effects of sample surface treatments on laser flash thermal diffusivity measurements.* Infrared physics & technology, 2002. **43**(3-5): p. 133-138.
17. Ozisik, M.N., *Inverse heat transfer: fundamentals and applications*. 2000: CRC Press.
18. Cheng, B., B. Lane, J. Whiting, and K. Chou, *A Combined Experimental-Numerical Method to Evaluate Powder Thermal Properties in Laser Powder Bed Fusion.* Proceddings of the ASME 2018 13th International Manufacturing Science and Engineering Conference, 2018, TX, USA, 2018.
19. Whiting, J., B. Lane, K. Chou, and B. Cheng, *Thermal property measurement methods and analysis for additive manufacturing solids and powders.* Proceedings of the 28th International Solid Freeform Fabrication Symposium. Austin, TX, USA, 2017.

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

20. Anam, M.A., J. Dilip, D. Pal, and B. Stucker. *Effect of scan pattern on the microstructural evolution of Inconel 625 during selective laser melting*. in *Proceedings of 25th Annual International Solid Freeform Fabrication Symposium*. 2014.
21. Zhang, S., *Design, analysis, and application of a cellular material/structure model for metal based additive manufacturing process.* 2017.
22. Capriccioli, A. and P. Frosi, *Multipurpose ANSYS FE procedure for welding processes simulation.* Fusion engineering and Design, 2009. **84**(2-6): p. 546-553.
23. *Inconel 625 material properties. Available from: http://www.hightempmetals.com/techdata/hitempInconel625data.php. Accessed October 10, 2017.*
24. Han, Q., R. Setchi, S.L. Evans, and C. Qiu, *Three-dimensional finite element thermal analysis in selective laser melting of Al-Al2O3 powder.*
25. Vora, H.D. and N.B. Dahotre, *Multiphysics theoretical evaluation of thermal stresses in laser machined structural alumina.* Lasers in Manufacturing and Materials Processing, 2015. **2**(1): p. 1-23.
26. Angela, J., Z. Radovan, and P. Frantisek, *Archives of Mechanical Technology and Materials.*
27. Cheng, B., S. Shrestha, and K. Chou, *Stress and deformation evaluations of scanning strategy effect in selective laser melting.* Additive Manufacturing, 2016. **12**: p. 240-251.
28. Yang, J., S. Sun, M. Brandt, and W. Yan, *Experimental investigation and 3D finite element prediction of the heat affected zone during laser assisted machining of Ti6Al4V alloy.* Journal of Materials Processing Technology, 2010. **210**(15): p. 2215-2222.

Zhang, Shanshan; Lane, Brandon; Whiting, Justin; Chou, Kevin.
"An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

Solid Freeform Fabrication 2018: Proceedings of the 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference
Reviewed Paper

QUANTIFYING UNCERTAINTY IN LASER POWDER BED FUSION ADDITIVE MANUFACTURING MODELS AND SIMULATIONS

Tesfaye Moges[1], Wentao Yan[1], Stephen Lin[2], Gaurav Ameta[1], Jason Fox[1], and Paul Witherell[1]

[1]National Institute of Standards and Technology, Gaithersburg, MD 20899

[2]Northwestern University, Evanston, IL 60208

Abstract

Various sources of uncertainty that can potentially cause variability in the product quality exist at different stages of the laser powder bed fusion (L-PBF) process. To implement computational models and simulations for quality control and process optimization, quantitative representation of their predictive accuracy is required. In this study, a methodology to estimate uncertainties in L-PBF models and simulations is presented. The sources of uncertainty, including those due to modeling assumptions, numerical approximation, input parameters, and measurement error, are discussed in detail and quantified for low and high-fidelity melt pool simulation models. A design of experiments (DOE) approach is leveraged to quantify uncertainty due to input parameters and investigate their effects on output quantities of interest (QoIs). The result of this work is essential for understanding the tradeoffs in model fidelity and guiding the selection of a model suitable for its intended purpose.

Keywords: Additive manufacturing, Powder bed fusion, Uncertainty quantification, Melt pool model, Design of experiments

Introduction

Metal additive manufacturing (AM) builds metallic parts layer upon layer directly from a 3D model. Due to its capability of producing metallic components having complex geometry and internal structures, it has become popular in different sectors, such as aerospace and automotive (scaffold cooling and weight reduction) and biomedical implants (prosthesis and femur structures) [1–3]. Laser powder bed fusion (L-PBF) is the most widely used metal AM process. In L-PBF, a thin layer of powder material is selectively scanned using a laser as the energy source to fuse powder particles together. In this process, powder layer formation and laser scanning operations are repeated multiple times until the final part is produced [4]. Due to the inherent nature of the L-PBF process, multiple physical phenomena and process parameters are involved at different stages. The process is governed by a variety of physical mechanisms, such as powder layer formation, laser-powder particle interaction, heat transfer, fluid dynamics of the melt pool, phase transformations, and microstructure evolution. Any variation encountered in these mechanisms potentially affects the quality of the final part. A major barrier that hinders full adoption of the L-PBF technology is inconsistent product quality.

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

To improve the quality of the product, it is important to investigate the sources of variability, quantify the uncertainties that occur at different stages of the process, and determine the magnitude of their effects on output quantities of interest. To quantify the uncertainties and analyze the sensitivity of input parameters, extensive research efforts have been continued based on physical experiments and computational models and simulations. Since performing uncertainty quantification in the L-PBF process using physical experiments is often expensive, computational models and simulations are promising tools to understand the dynamics, complex phenomena, and variabilities existing in the process. Although promising, computational models and simulations have not been fully utilized in quality control and process optimization due to lack of quantitative representation of their predictive accuracy. Without knowledge of the degree of accuracy of the L-PBF models and simulations, it is challenging to select a suitable model for the intended purpose. Therefore, it is necessary to identify and quantify the potential sources of uncertainty to investigate the predictive accuracy of such models and simulations.

This paper presents the sources of uncertainty in computational models and in the experimental validation and a methodology to quantify the uncertainties that exist in the L-PBF models. The predictive accuracy of such models strongly depends on the included and neglected physics of the process. *Modeling uncertainty* originates from the modeling assumptions that neglect part of the physical phenomena of a process. In addition, computational models require several input parameters including process parameters and material properties to represent the physical scenario of the process. However, the value of some parameters cannot always be known precisely and may exhibit inherent temporal fluctuations. Therefore, there is an associated *parameter uncertainty* in the computational models due to unknown input parameters. Moreover, the mathematical equations used to formulate the physical phenomena are difficult to be solved analytically, and various numerical methods have been used to discretize the system into finite elements and temporal transient phenomena into time steps to obtain an approximate solution. This discretization introduces *numerical uncertainty* in the computational models. Lastly, to validate the simulation results against measurement data, experimental results introduce *measurement uncertainty* due to imprecise measurement methods. In this study, all of these sources of uncertainty are quantified for the Rosenthal's semi-analytical thermal model [5,6]. In addition, we recommend best practices for quantifying model uncertainty for a finite element method (FEM)-based thermal model [7].

In this paper, we first briefly present the state-of-the-art in uncertainty quantification for L-PBF models. We then explore the shortcomings of previous work by classifying and characterizing sources of uncertainty in L-PBF models. Then, we examine, model, and evaluate all uncertainty sources of a L-PBF model using a case study through a standards-based approach. We view this work to be an essential step to report on requirements for further standardization and ultimately guide the selection of models suitable for their intended purpose.

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

Background

Since experimental-based part qualification for the L-PBF process is time consuming and costly, model-based qualification has been the focus of much research in the AM community. Several computational models have been developed to simulate the physical mechanisms and predict output quantities of interest at different stages of the process [8,9]. These models can be classified as powder bed models [10,11], heat source models [12,13], melt pool models [14,15], solidification models [16,17], and residual stress and distortion models [18,19]. Though computational models are different in their formulation, they share similar abstraction and characteristics [20,21]. The fidelity of the L-PBF models can be evaluated by identifying the various sources of uncertainty and quantifying their individual contribution to the overall prediction uncertainty [20].

Uncertainty quantification (UQ) of AM processes has recently been receiving increasing attention and some research efforts exist based on physical experiments, modeling and simulations [9,22,23]. Several reports have focused on experimentally investigating the effects of process parameters and material properties on the output quantities of interest and performing uncertainty quantification and sensitivity analysis [24–27]. Since AM processes possess a large number of parameters that influence the quality of a product, experimental-based UQ is expensive [28]. Thus, model-based UQ is getting attention for part qualification. Moser et al. [29] used a stochastic collocation approach to identify the most important parameters that affect the fidelity of FEM thermal model. Assuming probability distribution functions (PDFs) of input parameters, such as powder density, thermal conductivity, specific heat, particle diameter, and simulation time, the PDF of peak temperature is predicted. Ma et al. [30] used fractional factorial DOE to identify the critical process parameters and material properties that significantly influence peak temperature in FEM-based thermal model. Nath et al. [31] conducted uncertainty analysis on a FEM-based melt pool model that determined temperature profile and investigated uncertainty propagation to the solidification model to quantify the uncertainty in the grain size distribution of the microstructure. There have also been research efforts on implementing UQ methods for powder-scale AM model by developing surrogate models [32].

Previous studies have mainly focused on investigating the effects of input parameters on output quantities of interest and only quantifying parameter uncertainty. However, computational models exhibit all of the abovementioned sources of uncertainty and UQ-based study should include the remaining uncertainty sources to accurately determine the fidelity of a model. Recently, Lopez et al. [5] identified the four sources of uncertainty and conceptualized the quantification approach on the Isotherm Migration Method (IMM) model [6] by choosing melt pool width as the output quantity of interest. We further extend UQ approaches in the present study to quantify *all* sources of uncertainty and analyze their contribution towards model fidelity considering semi-analytical and FEM-based L-PBF melt pool models as a case study.

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

Uncertainty Quantification of Computational Models

In this section, we present a detailed discussion of the sources of uncertainty and their quantification, including those due to modeling assumptions, numerical approximation, input parameters, and measurement error of L-PBF process models.

Modeling Uncertainty

Computational models do not exactly represent the physical mechanisms that exists in L-PBF process as they are developed based on assumptions that neglect or simplify some phenomena. Thus, there is always a discrepancy between simulation results and true physical mechanisms. Modeling uncertainty originates from assumptions and simplifications made in computational models. There are a number of predictive models in the literature to represent the same L-PBF physical process [8,9]. These models are developed based on different assumptions and, therefore, there may be a significant discrepancy in their predictive accuracy due to modeling uncertainty. For example, to determine the packing density of the powder bed, powder bed models have been developed based on the Raindrop algorithm or the discrete element method. These models can induce different results and have different fidelity due to the different assumptions considered including those for powder particles shape, size and distribution, inter-particles forces, and boundary conditions. Similarly, there are various melt pool models in the literature to determine the temperature field and melt pool dimensions in the L-PBF process. These models are developed based on Rosenthal's semi-analytical thermal model, on a FEM thermal model, on a lattice Boltzmann approach, or on computational fluid dynamics (CFD). These models have different assumptions in terms of considering powder bed as a continuum or as particles, energy source as a point or distributed, and absorbed energy as a surface or volumetric distribution. There are also differences in considering other physical phenomena, such as surface tension, the Marangoni effect, recoil pressure, vaporization, capillary, and wetting. Although these models are developed for predicting the same output quantities of interest, their modeling uncertainty results in different predictive accuracy due to the different assumptions they considered. Therefore, it is important to quantify modeling uncertainty of computational models to determine their degree of predictive accuracy.

To quantify modeling uncertainty, simulation results $\boldsymbol{S}$ of the predictive model need to be validated against the experimental data $\boldsymbol{D}$. The ASME V&V-20 standard [33], which discusses the sources of uncertainty and UQ methods in heat transfer and fluid mechanics models, can be suitable for L-PBF models as it involves thermally-activated consolidation processes [5]. The interval within which modeling error falls is characterized by $\boldsymbol{\delta_{model}} \boldsymbol{\epsilon} [\boldsymbol{E} \pm \boldsymbol{u_{val}}]$ where, $\boldsymbol{E}$ is the comparison error between simulation result $\boldsymbol{S}$ and measurement data $\boldsymbol{D}$, and $\boldsymbol{u_{val}}$ is validation uncertainty that accounts for all sources of uncertainty. Assuming that they are independent, it can be computed as:

$$u_{val} = \sqrt{u_{num}^2 + u_{input}^2 + u_D^2}, \quad \text{.....(1)}$$

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

where u_{num}, u_{input}, and u_D are numerical uncertainty, parameter uncertainty, and measurement uncertainty, respectively. The following sub-sections discuss the uncertainty sources and quantification methods of these uncertainties.

Numerical Uncertainty

Due to the complexity of the L-PBF process, constitutive equations that approximate the physical phenomena are not often solved using analytical methods. Numerical methods that discretize the time and length variables are used to solve the partial differential equations. For the L-PBF process, predictive models are commonly developed based on numerical methods, such as finite element models, discrete element models, lattice Boltzmann method, and computational fluid dynamics studies. Thus, the choice of finite time and length resolution introduces numerical error that undermines the accuracy of the simulation results of the output quantities of interest [34]. For example, the element size or number of elements in an FEM-based thermal model and mesh discretization to represent the change in temperature in semi-analytical thermal model cause numerical uncertainty in the predicted melt pool width. Since most of predictive models are computationally expensive, reduced order and surrogate models are used to statistically represent the simulation models. This can introduce additional uncertainty due to the limited number of training data used to build the representative model.

Numerical uncertainty can be quantified using a grid convergence index (GCI) developed by Roache [35]. The GCI is an error percentage that provides an estimate of the coverage interval within which the numerical error will likely lie. Numerical uncertainty is the GCI percentage of the value of the output quantity of interest. The GCI is obtained by multiplying the absolute value of Richardson extrapolation error by a safety factor determined through empirical studies and given by:

$$GCI = F_s \frac{\epsilon_{ext}}{r_{21}^p - 1}, \quad \text{.....(2)}$$

$$\epsilon_{ext} = \left| \frac{f_{ext}^{21} - f_1}{f_{ext}^{21}} \right|, \quad f_{ext}^{21} = \frac{r_{21}^p f_1 - f_2}{r_{21}^p - 1}, \quad p = \frac{\ln(|f_3 - f_2|/|f_2 - f_1|) + q(p)}{ln r_{21}}, \quad \text{.....(3)}$$

where, $F_s = 1.25$ is factor of safety; f_1, f_3, and f_3 are three simulation results at fine h_1, finer h_2, and finest h_3 mesh sizes, respectively; and $r_{21} = h_2/h_1$ is the mesh refinement ratio. The order of convergence p is determined using Equation (3). For a constant mesh refinement ratio, $q(p) = 0$. Otherwise, $q(p) = ln\left[\left(r_{21}^p - s\right)/\left(r_{32}^p - s\right)\right]$ and $s = sign[(f_3 - f_2)/(f_2 - f_1)]$, and the convergence order is solved iteratively with initial guess $q(p) = 0$.

Parameter Uncertainty

Computational models require prior determination of input parameters to represent the behavior of a process. The values of some parameters are not precisely known due to natural variation in the system or lack of sufficient data and knowledge to determine the exact value. Therefore, the assigned value has uncertainty that propagates into an output quantity of interest.

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

The sources of uncertainty can take one of two forms: aleatory and epistemic. Aleatory uncertainty arises due to natural variation existing in the parameters and in the performance of the system. For instance, inherent drift and fluctuation in the laser and galvanometer systems cause aleatory variability in the laser power and scan speed in L-PBF process. On the other hand, epistemic uncertainty arises as a result of lack of knowledge regarding the behavior of a system. For example, the determination of the absorption coefficient of a powder bed is not well established due to lack of a convincing measurement method and could be resolved by introducing additional information. This study considers the joint effect of aleatory and epistemic uncertainties instead of distinguishing them separately as suggested by Roache [35].

To quantify the input parameter uncertainty, first the uncertainty associated with the parameters is captured in a form of distribution, nominal value, and standard deviation. Then, the sources of uncertainty of these parameters propagate into output quantity of interest through computational models or reduced order models. In this study, a design of experiments (DOE) method is used to quantify input parameter uncertainty by formulating a reduced order formulation. Experimental design is a suitable technique to identify the most important factors that have significant impact on the response variable and develop a response surface model that approximates the original process [36]. In the present study, a fractional factorial design of experiments approach is used to plan the design matrix for simulation runs and quantify input parameter uncertainty, choosing melt pool width as a response variable. The detailed procedure and discussion on implementing DOE for UQ of input parameters is presented in the case study section.

Measurement Uncertainty

Since experimental data is required to validate the simulation results, measurement uncertainty is important in the UQ process to determine the predictive accuracy of L-PBF models. Measurement uncertainty mainly depends on the methods and equipment used for data acquisition. To understand the process-structure-properties-performance chain of the L-PBF process, measurement results during pre-process, in-process, and post-process are necessary [37]. For example, uncertainties related to measurement methods used to determine the powder packing structure that depends on metal powder characterization (such as powder size, morphology, density, and distribution [38,39]), non-intrusive infrared thermography, and pyrometry to measure surface temperature of the heat affected zone [37]. Measurement uncertainty is quantified as per the guide to the expression of uncertainty in measurement (GUM) that standardized the evaluation and expression of uncertainty in measurement [40].

Propagation of Uncertainty in L-PBF Models

Uncertainty is propagated through L-PBF models in the manner depicted in Figure 1. Input uncertainty enters each different model with input parameters. Outputs of some models are incorporated as inputs in other models. For this reason, it is required to understand the propagated effects of uncertainty through a composition of models. Besides parameter uncertainty, modeling and numerical uncertainties also propagate to outputs.

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

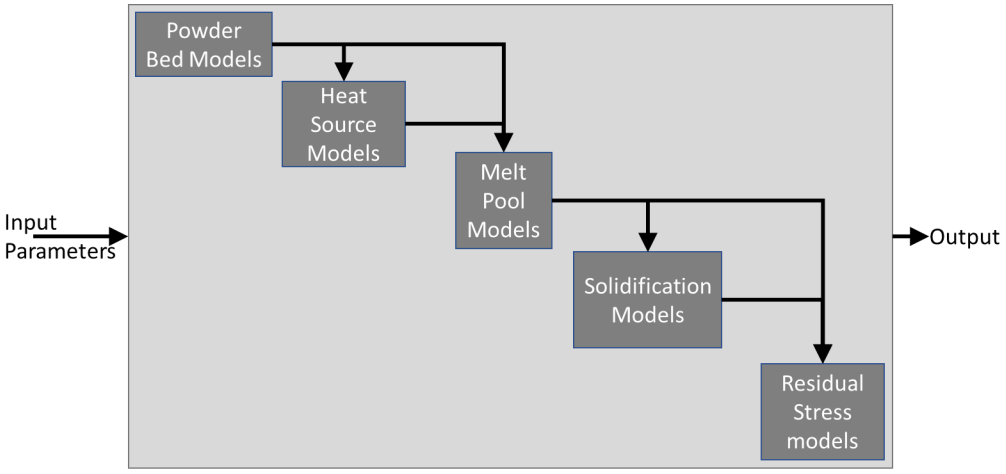


Figure 1: Flow of uncertainty in L-PBF models.

Case Study: UQ in a Semi-Analytical- and FEM-Based Melt Pool Models

In this case study, semi-analytic- and FEM-based melt pool models are selected for the purpose of quantifying all sources of uncertainty in the L-PBF process as shown in the previous section. In a semi-analytical-based melt pool model, the heat conduction equation is transformed into a set of ordinary differential equations to determine the isotherm velocities on the surface of the powder bed in terms of positions and temperature derivatives. The phase change that occurs during the process is taken into consideration to incorporate the effect of the internal energy difference between solid and liquid states at melting temperature, which is given by the latent heat of fusion of a material. The melt pool width is predicted directly from the isotherm position by assigning a melting temperature on one of the isotherms. The model is first developed for laser cladding [6] and adjusted for prediction of melt pool dimensions in L-PBF [5]. Although this model considers the temperature-dependent material properties and provides results in an efficient manner, there are a number of assumptions related to the phenomena of the process. The heat source is assumed as a point source, which is not in conjunction with reality instead of a distributed one. The melt pool flow and distribution of particles in a powder bed are also ignored. This simplification and assumptions are expected to increase modeling error. The discretization of the temperature increment to represent the isotherms creates numerical error.

The FEM-based melt pool model is the most popular method to simulate the L-PBF process [8]. It discretizes the powder bed into a finite number of elements by forming a mesh to solve the partial differential equation governing the system in order to estimate thermal field and melt pool characteristics. The main assumption of this model is that the powder bed is considered a homogenous continuum material instead of a distribution of powder particles. The physical phenomena encountered in the melt pool dynamics, such as surface tension and the Marangoni effect, are ignored, which creates modeling error. The FEM-based melt pool model used for present work is proposed by Smith et al. [7] assuming a heat source with Gaussian distribution.

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

Computational Design of Experiments

To make the uncertainty quantification process for a model having many parameters more computationally practical, a DOE approach, which is widely used to identify the most important parameters that have major influence on the output, is used for the present work. As the number of factors increases, investigating the effect of each of the factors along with their interactions on the output quantities of interest using full factorial DOE is infeasible due to high computational cost. Therefore, a fractional factorial DOE is chosen to sample some of the most important runs that can provide the necessary information about the main effects and second-order interactions [41].

For the semi-analytical-based melt pool model, all the input parameters (i.e. nine factors) used in the model are selected for DOE analysis. A 2_{IV}^{9-4} fractional factorial design that represents two levels for each factor and four resolution is selected. In this design, no main effects are confounded with any other main effect and two-factor interactions. We select 2_{IV}^{10-5} fractional factorial design having ten factors for the FEM-based melt pool model. These designs require $2^5 = 32$ simulation runs for each model. Normal distributions, which commonly used in UQ to represent the variations of input parameters due to random and imperfect knowledge, are assumed for the selected factors and the design matrices for the semi-analytical- and FEM-based melt pool models are outlined in Tables 1 and 2, respectively.

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

Table 1: DOE plan for the semi-analytical-based melt pool model

Factor	X1	X2	X3	X4	X5	X6	X7	X8	X9
Name	Laser power	Scan speed	Preheat temper-ature	Density	Specific heat capacity	Thermal conduct-ivity	Absorption coefficient	Latent heat of fusion	Melting temper-ature
Symbol	P	v	T_o	ρ	c_p	k	A	h_l	T_m
Nominal value	195	0.8	293	$\rho(T)$	$c_p(T)$	$k(T)$	0.4	$2.97\text{x}10^5$	1593
Unit	W	m/s	K	kg/m^3	J/kgK	W/mK		J/kg	K
Std. dev.	2.5%	1.5%	1%	1%	3%	3%	25%	5%	5%
Run 01	−	−	−	−	−	+	+	+	+
Run 02	+	−	−	−	−	+	−	−	−
Run 03	−	+	−	−	−	−	+	−	−
Run 04	+	+	−	−	−	−	−	+	+
Run 05	−	−	+	−	−	−	−	+	−
Run 06	+	−	+	−	−	−	+	−	+
Run 07	−	+	+	−	−	+	−	−	+
Run 08	+	+	+	−	−	+	+	+	−
Run 09	−	−	−	+	−	−	−	−	+
Run 10	+	−	−	+	−	−	+	+	−
Run 11	−	+	−	+	−	+	−	+	−
Run 12	+	+	−	+	−	+	+	−	+
Run 13	−	−	+	+	−	+	+	−	−
Run 14	+	−	+	+	−	+	−	+	+
Run 15	−	+	+	+	−	−	+	+	+
Run 16	+	+	+	+	−	−	−	−	−
Run 17	−	−	−	−	+	−	−	−	−
Run 18	+	−	−	−	+	−	+	+	+
Run 19	−	+	−	−	+	+	−	+	+
Run 20	+	+	−	−	+	+	+	−	−
Run 21	−	−	+	−	+	+	+	−	+
Run 22	+	−	+	−	+	+	−	+	−
Run 23	−	+	+	−	+	−	+	+	−
Run 24	+	+	+	−	+	−	−	−	+
Run 25	−	−	−	+	+	+	+	+	−
Run 26	+	−	−	+	+	+	−	−	+
Run 27	−	+	−	+	+	−	+	−	+
Run 28	+	+	−	+	+	−	−	+	−
Run 29	−	−	+	+	+	−	−	+	+
Run 30	+	−	+	+	+	−	+	−	−
Run 31	−	+	+	+	+	+	−	−	−
Run 32	+	+	+	+	+	+	+	+	+

The design matrices in Tables 1 and 2 are coded as (–) and (+) to represent the low (nominal value minus standard deviation) and high (nominal value plus standard deviation) values of the two levels for each of the factors and defines the 32 simulation runs. The matrices in Tables 1 and 2 are arranged in such a way that all columns are orthogonal to each other and the main and interaction effects can be independently estimated. Since the variation of input parameters and some of their values are not yet explicitly determined, the nominal values and their variations are chosen based on prior research and expert opinions [30].

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

Table 2: DOE plan for the FEM-based melt pool model

Factor	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Variable name	Laser power	Scan speed	Layer thick-ness	Laser beam radius	Density	Specific heat capacity	Thermal conduct-ivity	Absorption coefficient	Latent heat of fusion	Emis-sivity
Symbol	P	v	l_t	r_{beam}	ρ	c_p	k	A	h_l	ε
Nominal value	195	0.8	40	45	$\rho(T)$	$c_p(T)$	$k(T)$	0.4	$2.97\text{x}10^5$	0.4
Units	W	m/s	μm	μm	kg/m^3	J/kgK	W/mK		J/kg	
Std. dev.	2.5%	1.5%	25%	10%	1%	3%	3%	25%	5%	10%
Run 01	−	−	−	−	−	+	+	+	+	+
Run 02	+	−	−	−	−	−	−	−	−	+
Run 03	−	+	−	−	−	−	−	−	+	−
Run 04	+	+	−	−	−	+	+	+	−	−
Run 05	−	−	+	−	−	−	−	+	−	−
Run 06	+	−	+	−	−	+	+	−	+	−
Run 07	−	+	+	−	−	+	+	−	−	+
Run 08	+	+	+	−	−	−	−	+	+	+
Run 09	−	−	−	+	−	−	+	−	−	−
Run 10	+	−	−	+	−	+	−	+	+	−
Run 11	−	+	−	+	−	+	−	+	−	+
Run 12	+	+	−	+	−	−	+	−	+	+
Run 13	−	−	+	+	−	+	−	−	+	+
Run 14	+	−	+	+	−	−	+	+	−	+
Run 15	−	+	+	+	−	−	+	+	+	−
Run 16	+	+	+	+	−	+	−	−	−	−
Run 17	−	−	−	−	+	+	−	−	−	−
Run 18	+	−	−	−	+	−	+	+	+	−
Run 19	−	+	−	−	+	−	+	+	−	+
Run 20	+	+	−	−	+	+	−	−	+	+
Run 21	−	−	+	−	+	−	+	−	+	+
Run 22	+	−	+	−	+	+	−	+	−	+
Run 23	−	+	+	−	+	+	−	+	+	−
Run 24	+	+	+	−	+	−	+	−	−	−
Run 25	−	−	−	+	+	−	−	+	+	+
Run 26	+	−	−	+	+	+	+	−	−	+
Run 27	−	+	−	+	+	+	+	−	+	−
Run 28	+	+	−	+	+	−	−	+	−	−
Run 29	−	−	+	+	+	+	+	+	−	−
Run 30	+	−	+	+	+	−	−	−	+	−
Run 31	−	+	+	+	+	−	−	−	−	+
Run 32	+	+	+	+	+	+	+	+	+	+

Results and Discussion

To identify the input parameters that significantly influence the melt pool width, the main and interaction effects are computed. An effect is the amount of change in melt pool width when only the parameter under consideration is changed from its low (–) to high (+) level. A normal probability plot of the effects is used to isolate statistically significant effects of factors and their interactions from those effects that come solely from random variables. The normal probability plot is used for assessing whether or not a data set is approximately normally distributed and identifying statistically significant factors [42]. The normal probability plot of the effects for the semi-analytical-based melt pool model is shown in Figure 2. The main and interaction effects that represent x-axis in Figure 2 are calculated by subtracting the average of the response at the low level from the high level for the parameter under consideration.

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

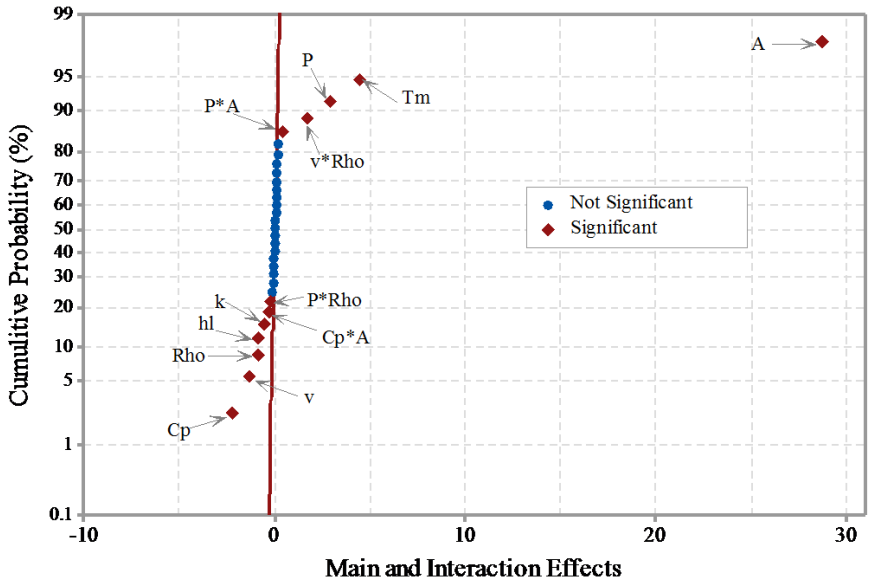


Figure 2: Normal probability plot of the semi-analytical-based melt pool model

It can be seen from normal probability plot that the main factors and some of their interactions significantly affect the melt pool width. A statistically-driven mathematical model is formulated from the DOE analysis using the identified main factors and interactions that have significant effect on the response [36]. At the given set of parameters, the mathematical model closely approximates the physics-based model. To quantify the input parameters uncertainty, a Monte Carlo approximation is conducted for 50,000 samples which are quite sufficient to get stable results and the probability distribution of the predicted melt pool width is shown in Figure 3. The average and standard deviation of the predicted melt pool width are found to be 94.2µm and 55.1µm, respectively.

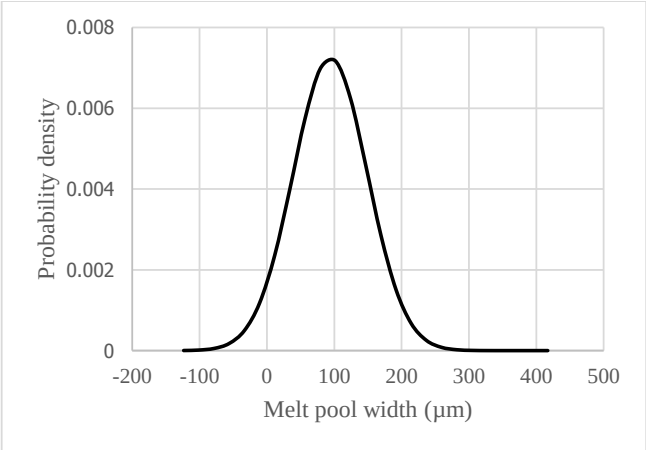


Figure 3: Normal distribution of predicted melt pool width of the semi-analytical model

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

The FEM-based melt pool model, though more accurate, posed difficulty when managing uncertainty with the calibration parameters. Using the set of input parameters in Table 2 the parameters that significantly affect melt pool width cannot be identified from the main and interaction effects due to the large percentage variation assumed for layer thickness and absorption coefficient. Thus, the uncertainty of the input parameters in which the model performs reasonably needed to be reconsidered.

To quantify numerical uncertainty in the semi-analytical melt pool model, simulations with a different number of isotherms having different temperature increments were run. In this case, all the input parameters are set to their nominal values. Then, the grid convergence index (GCI), which is an estimate of 95% uncertainty, was computed using Equation (2). Using the given nominal values, the estimated melt pool width in the semi-analytical model is 100.9 ± 3.43μm. Similarly, a convergence study was performed for the FEM-based melt pool model by running the simulation with different element sizes. Using the given nominal values, the estimated melt pool width for the FEM-based model is 141.0 ± 4.23μm.

To complete UQ in L-PBF models, measurement uncertainty that comes from the experiments used for validation is required. The melt pool width was measured from the image of a 1mm long scan track of IN625 captured using an optical microscope by manually tracing the edges of the track and determining the distance between the traces. The average and standard deviation of the melt pool width at 195W and 800mm/s are measured to be 132.2μm and 14.1μm, and the 95% suggested confidence interval is ± 28.2μm. In addition to this uncertainty, manually tracing the edges of the scan track causes uncertainty due to human error estimated to be ± 2μm. This is based on the ability to determine the edge of the track accurately accounting for the focus of the image at the edge of the track and the size of pixels. More details about the experimental results and measurement methodology is given by Fox et al. [43].

The comparison of the predicted and measured melt pool width by the two models at a given laser power and scan speed combination is shown in Figure 4. Assuming all sources of uncertainty are independent, the validation uncertainty that accounts for all sources of uncertainty is computed using Equation (1) and for the semi-analytical model is estimated to be ± 114.3μm. Thus, the interval within which the modeling error falls for the nominal conditions is 31.3 ± 114.3μm. Since the uncertainty of the semi-analytical melt pool model at the given set of parameters is large and beyond the realm of possibility, the percentage variation assumed for absorption coefficient need to be revised. From the DOE analysis and obtained results, the following observations can be made.

(1) The modeling uncertainty of the semi-analytical-based melt pool model ($E = 31.3$μm and $u_{val} = 114.3$μm at nominal parameter setup) is large as expected due to many assumptions and simplifications used regarding heat source, powder layer formation, and melt pool dynamics. Ignoring these phenomena results in larger modeling error,

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

and thus a model that considers many physical phenomena existing in L-PBF process can have better predictive accuracy.

(2) The contribution of input parameters uncertainty to the overall modeling uncertainty is more than 100% (± 109.2%) for the semi-analytical model. This is mainly due to the large uncertainty value assumed for the input parameters, especially, the absorption coefficient (± 25%). The knowledge of parameter uncertainty is necessary for the estimation of modeling uncertainty in which the model accurately determines the response within a specified range of parameters. In addition, the accuracy of the statistically-driven mathematical model can be improved by increasing the number of testing results which requires more simulation runs.

(3) The contribution of the numerical uncertainty is very small (± 3.4% for semi-analytical) compared to the other sources of uncertainty. This contribution can be considered negligible. The measurement uncertainty used for model validation has a significant contribution to the modeling uncertainty (± 29.9%). This large variation can be associated with the calibration of the optical microscope used to capture the image of a scan track, variation of the incandescent light generated by the hot surface (which depends on the temperature and emissivity of the surface), the adjusted exposure time of the camera, and the dynamics of the melt pool, powder, and laser interactions [43].

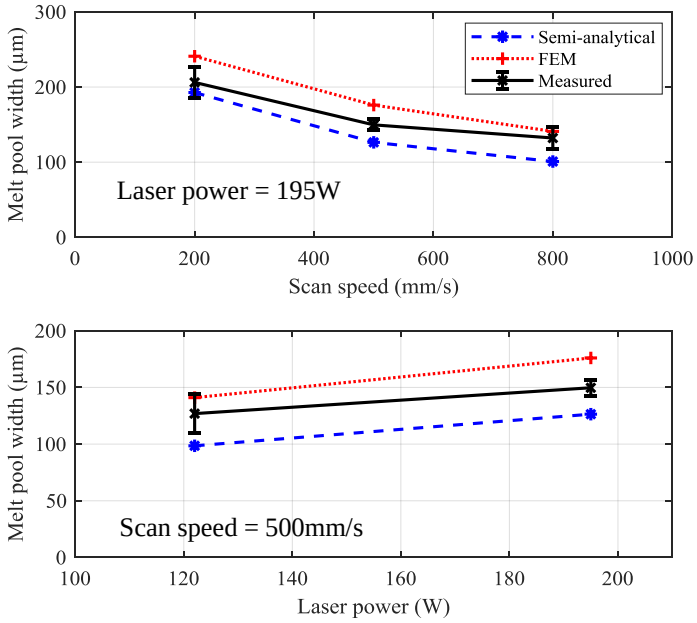


Figure 4: Predicted and measured melt pool width

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

Conclusions

In this paper, we presented an uncertainty quantification strategy for L-PBF models. A case study was presented with two models: a semi-analytical-based and a FEM melt pool models. A DOE fractional factorial study was conducted with two levels for nine and ten input parameters for the semi-analytical-based and FEM melt pool models, respectively. The DOE models were then used as part of a Monte Carlo simulation to predict melt pool width in order to compute output uncertainty of the models due to uncertainty in input parameters.

The semi-analytical-based melt pool model was computationally efficient to run, but the DOE prediction results had large uncertainty. The FEM melt pool model, though more accurate, posed difficulty when managing uncertainty with the calibration parameters. The DOE study conducted in this paper included only two levels and therefore can only estimate linear effects between the levels. To develop an accurate DOE model for prediction, three or more levels would need to be conducted. This requires an increasingly large number of simulations for quantifying the uncertainty of FEM models.

Future work will involve further refining the amount of uncertainty included in the input parameters and then running the DOE with higher levels that ensure non-linear effects between the levels and among multiple factors. This process will be evaluated for both the semi-analytical-based and FEM-based melt pool models. Refinement of the input uncertainty especially on absorption coefficient and layer thickness is needed as the amount of uncertainty assumed for these parameters is appeared to be too high.

Disclaimer

No approval or endorsement of any commercial product by NIST is intended or implied. Certain commercial equipment, instruments or materials are identified in this report to facilitate better understanding. Such identification does not imply recommendations or endorsement by NIST nor does it imply the materials or equipment identified are necessarily the best available for the purpose.

Acknowledgments

The authors gratefully acknowledge Yan Lu, Brandon Lane, Moneer Helu, William Bernstein, and Thomas Kramer from NIST for their valuable feedback that improved the paper.

References

[1] Petrovic, V., Vicente Haro Gonzalez, J., Jordá Ferrando, O., Delgado Gordillo, J., Ramón Blasco Puchades, J., and Portolés Griñan, L., 2011, "Additive Layered Manufacturing: Sectors of Industrial Application Shown through Case Studies," Int. J. Prod. Res., 49(4), pp. 1061–1079.

[2] Wohlers, T., 2016, "Wohlers Report 2016. 3D Printing and Additive Manufacturing State of the Industry," Wohlers Rep. 2016, (May), p. 355.

[3] Coykendall, J., Cotteleer, M., Holdowsky, L., and Mahto, M., 2014, "3D Opportunity in Aerospace

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

and Defense," Deloitte Univ. Press, pp. 1–28.
[4] Hodge, N. E., Ferencz, R. M., and Vignes, R. M., 2016, "Experimental Comparison of Residual Stresses for a Thermomechanical Model for the Simulation of Selective Laser Melting," Addit. Manuf., 12, pp. 159–168.
[5] Lopez, F., Witherell, P., and Lane, B., 2016, "Identifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models," J. Mech. Des., 138(November), pp. 1–4.
[6] Devesse, W., De Baere, D., and Guillaume, P., 2014, "The Isotherm Migration Method in Spherical Coordinates with a Moving Heat Source," Int. J. Heat Mass Transf., 75, pp. 726–735.
[7] Smith, J., Xiong, W., Cao, J., and Liu, W. K., 2016, "Thermodynamically Consistent Microstructure Prediction of Additively Manufactured Materials," Comput. Mech., 57(3), pp. 359–370.
[8] Schoinochoritis, B., Chantzis, D., and Salonitis, K., 2014, "Simulation of Metallic Powder Bed Additive Manufacturing Processes with the Finite Element Method: A Critical Review," Proc. Inst. Mech. Eng. Part B J. Eng. Manuf., 231(1), pp. 96–117.
[9] Hu, Z., and Mahadevan, S., 2017, "Uncertainty Quantification and Management in Additive Manufacturing: Current Status, Needs, and Opportunities," Int. J. Adv. Manuf. Technol., pp. 1–20.
[10] Zhou, J., Zhang, Y., and Chen, J. K., 2009, "Numerical Simulation of Random Packing of Spherical Particles for Powder-Based Additive Manufacturing," J. Manuf. Sci. Eng., 131(3), pp. 1-8.
[11] Herbold, E. B., Walton, O., and Homel, M. A., 2015, "Simulation of Powder Layer Deposition in Additive Manufacturing Processes Using the Discrete Element Method," LLNL-TR-678550.
[12] Boley, C. D., Khairallah, S. A., and Rubenchik, A. M., 2015, "Calculation of Laser Absorption by Metal Powders in Additive Manufacturing.," Appl. Opt., 54(9), pp. 2477–82.
[13] Gusarov, A. V., and Kruth, J. P., 2005, "Modelling of Radiation Transfer in Metallic Powders at Laser Treatment," Int. J. Heat Mass Transf., 48(16), pp. 3423–3434.
[14] Bo Cheng; Kevin Chou, 2015, "Melt Pool Evolution Study in Selective Laser Melting," Dyn. Syst. with Appl. using MATLAB, 53(August), pp. 1182–1194.
[15] Khairallah, S. A., Anderson, A. T., Rubenchik, A., and King, W. E., 2016, "Laser Powder-Bed Fusion Additive Manufacturing: Physics of Complex Melt Flow and Formation Mechanisms of Pores, Spatter, and Denudation Zones," Acta Mater., 108, pp. 36–45.
[16] Li, X., and Tan, W., 2017, "3-Dimensional Cellular Automata Simulation of Grain Structure in Metal Additive Manufacturing Processes," Solid Free. Fabr. 2017 Proc. 28th Annu. Int. Solid Free. Fabr. Symp. – An Addit. Manuf. Conf., pp. 1030–1047.
[17] Keller, T., Lindwall, G., Ghosh, S., Ma, L., Lane, B. M., Zhang, F., Kattner, U. R., Lass, E. A., Heigel, J. C., Idell, Y., Williams, M. E., Allen, A. J., Guyer, J. E., and Levine, L. E., 2017, "Application of Finite Element, Phase-Field, and CALPHAD-Based Methods to Additive Manufacturing of Ni-Based Superalloys," Acta Mater., 139, pp. 244–253.
[18] Zaeh, M. F., and Branner, G., 2010, "Investigations on Residual Stresses and Deformations in Selective Laser Melting," Prod. Eng., 4(1), pp. 35–45.
[19] Cheng, B., Shrestha, S., and Chou, K., 2016, "Stress and Deformation Evaluations of Scanning Strategy Effect in Selective Laser Melting," Addit. Manuf., 12, pp. 240–251.
[20] Lopez, F., Witherell, P., and Lane, B., 2016, "Identifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models," Proc. ASME 2016 Int. Manuf. Sci. Eng. Conf. MSEC2016 June 27-July 1, 2016, Blacksburg, Virginia, USA, 138(November), pp. 1–10.
[21] Assouroko, Ibrahim; Lopez, Felipe; Witherell, P., 2016, "A Method for Characterizing Model Fidelity in Laser Powder Bed Fusion Additive Manufacturing," Proc. ASME 2016 Int. Mech. Eng. Congr. Expo. ASME IMECE 2016 Novemb. 11-17, 2016, Phoenix, Arizona, USA, pp. 1–13.
[22] Tapia, G., and Elwany, A., 2014, "A Review on Process Monitoring and Control in Metal-Based Additive Manufacturing," J. Manuf. Sci. Eng., 136(6), pp. 060801-1-060801-10.
[23] Garg, A., Tai, K., and Savalani, M. M., 2014, "State-of-the-Art in Empirical Modelling of Rapid Prototyping Processes," Rapid Prototyp. J., 20(2), pp. 164–178.
[24] Partee, B., Hollister, S. J., and Das, S., 2006, "Selective Laser Sintering Process Optimization for Layered Manufacturing of CAPA 6501 Polycaprolactone Bone Tissue Engineering Scaffolds," J.

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

Manuf. Sci. Eng., 128(2), pp. 531–540.
[25] Adamczak, S., Bochnia, J., and Kaczmarska, B., 2014, "Estimating the Uncertainty of Tensile Strength Measurement for a Photocured Material Produced by Additive Manufacturing," Metrol. Meas. Syst., 21(3), pp. 553–560.
[26] Delgado, J., Ciurana, J., and Rodríguez, C. A., 2012, "Influence of Process Parameters on Part Quality and Mechanical Properties for DMLS and SLM with Iron-Based Materials," Int. J. Adv. Manuf. Technol., 60(5–8), pp. 601–610.
[27] Raghunath, N., and Pandey, P. M., 2007, "Improving Accuracy through Shrinkage Modelling by Using Taguchi Method in Selective Laser Sintering," Int. J. Mach. Tools Manuf., 47(6), pp. 985–995.
[28] Kamath, C., 2016, "Data Mining and Statistical Inference in Selective Laser Melting," Int. J. Adv. Manuf. Technol., 86(5–8), pp. 1659–1677.
[29] Moser, D., Beaman, J., Fish, S., and Murthy, J., 2014, "Multi-Layer Computational Modeling of Selective Laser Sintering Processes," Proc. ASME 2014 Int. Mech. Eng. Congr. Expo. IMECE2014, pp. 1–11.
[30] Ma, L., Fong, J., Lane, B., Moylan, S., Filliben, J., Heckert, A., and Levine, L., 2015, "Using Design of Experiments in Finite Element Modeling To Identify Critical Variables for Laser Powder Bed Fusion," Solid Free. Fabr. Symp., pp. 219–228.
[31] Nath, P., Hu, Z., and Mahadevan, S., 2017, "Multi-Level Uncertainty Quantification in Additive Manufacturing," Solid Free. Fabr. 2017 Proc. 28th Annu. Int. S, pp. 922–937.
[32] King, W. E., Anderson, A. T., Ferencz, R. M., Hodge, N. E., Kamath, C., Khairallah, S. A., and Rubenchik, A. M., 2015, "Laser Powder Bed Fusion Additive Manufacturing of Metals; Physics, Computational, and Materials Challenges," Appl. Phys. Rev., 2(4), p. 041304.
[33] ASME-V&V-20, 2009, An Overview of ASME V&V 20: Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer, American Society of Mechanical Engineers.
[34] Schwer, L. E., 2008, "Is Your Mesh Refined Enough? Estimating Discretization Error Using GCI," 7th LS-DYNA Anwenderforum, 1(1), pp. 45–54.
[35] Roache, P., 2002, "Code Verification by the Method of Manufactured Solutions," ASME J. Fluids Engineer., 114(1), pp. 4-10.
[36] Montgomery, D. C., 2012, *Design and Analysis of Experiments*, John Wiley & Sons, Inc.
[37] Mahesh, M., Lane, B., Donmez, A., Feng, S., Moylan, S., and Fesperman, R., 2015, "Measurement Science Needs for Real-Time Control of Additive Manufacturing Powder Bed Fusion Processes," Natl. Inst. Stand. Technol., pp. 1–50.
[38] Slotwinski, J. A., Garboczi, E. J., Stutzman, P. E., Ferraris, C. F., Watson, S. S., and Peltz, M. A., 2014, "Characterization of Metal Powders Used for Additive Manufacturing," J. Res. Natl. Inst. Stand. Technol., 119, pp. 460–493.
[39] Cooke, A., and Slotwinski, J., 2015, "Properties of Metal Powders for Additive Manufacturing: A Review of the State of the Art of Metal Powder Property Testing," Addit. Manuf. Mater. Stand. Test. Appl., pp. 21–48.
[40] JCGM, 2008, "Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement," Int. Organ. Stand. Geneva ISBN, 50(September), p. 134.
[41] Hassanpour, R. M., 2010, "Application of Experimental Design in Uncertainty Quantification," Pap. CCG Annu. Rep., 126(12), pp. 1–6.
[42] NIST/SEMATECH e–Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, [last accessed: August 31, 2018].
[43] Fox, J. C., Lane, B. M., and Yeung, H., 2017, "Measurement of Process Dynamics through Coaxially Aligned High Speed Near-Infrared Imaging in Laser Powder Bed Fusion Additive Manufacturing," Proc. SPIE 10214, Thermosense: Thermal Infrared Applications XXXIX 1(301), pp. 1021407

Moges, Tesfaye; Yan, Wentao; Lin, Stephen; Ameta, Gaurav; Fox, Jason; Witherell, Paul.
"Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations."
Paper presented at 29th Annual International Solid Freeform Fabrication Symposium: An Additive Manufacturing Conference, Austin, TX, United States. August 13, 2018 - August 15, 2018.

Development of a Detection Algorithm for Kitchen Cooktop Ignition Prevention

Amy Mensch
National Institute of Standards & Technology
100 Bureau Dr., MS-8664
Gaithersburg, MD 20899
301-975-6714
amy.mensch@nist.gov

Anthony Hamins
National Institute of Standards & Technology
100 Bureau Dr., MS-8664
Gaithersburg, MD 20899
301-975-6598
anthony.hamins@nist.gov

Kathryn Markell
National Institute of Standards & Technology
100 Bureau Dr., MS-8664
Gaithersburg, MD 20899

Background and Objectives

According to a recent NFPA report, 47 % of reported home fires involve cooking equipment, with cooktops accounting for 87 % of cooking-fire deaths and 80 % of the civilian injuries [1]. Electric-coil stovetops manufactured after June 2018 in the U.S. must pass the UL 858 [2] "abnormal cooking test." The test prescribes a maximum dry-pan temperature or an ignition-prevention performance test using 50 g of canola oil with the coil element on its highest power setting. This standard does not apply to older cooktops or other types of cooktops, such as gas ranges. Therefore, a set of experiments was designed to investigate the feasibility of a robust and reliable, external, pre-ignition detection system that could be used to retrofit existing cooktops. The goal of such a system would be to prevent fires from unattended cooking, while ignoring normal cooking activities and nuisance sources. The proposed system could be located within the kitchen exhaust duct or on the ceiling in the kitchen. It could also be integrated into existing household systems via the "internet of things."

There have been many studies investigating the performance advantages of multiple sensors over a single sensor for detection of generalized fire conditions and nuisance-alarm resistance. Gottuk et al. [3] compared the effectiveness of various multi-criteria fire-detection algorithms, using signals from carbon monoxide (CO) sensors and smoke detectors to reduce false fire alarms and to increase detection sensitivity. A cutoff value for the product of the signals from an ionization smoke detector and a CO sensor was reported to show improved effectiveness over typical smoke detectors.

In another study, Cestari et al. [4] included a thermocouple and ionization, photoelectric and CO detectors to 1) develop advanced fire detection algorithms that reduced nuisance sensitivity and 2) detect fires at least as fast as conventional ionization and photoelectric detectors. Eight parameters were identified from the four sensors by considering the magnitude and rate of rise of the output from each sensor. Algorithms developed from these parameters showed that the best fire sensitivity and nuisance immunity was observed for the algorithms based on: temperature rise and CO; CO and ionization detector; and temperature rise, CO and ionization detector. Another series of studies developed and tested a prototype four-sensor (ionization, photoelectric, CO and carbon dioxide (CO_2)) package for early warning seaboard applications [5]. Although these studies did not focus solely on cooktop fires, typical cooktop nuisance sources were considered, including steam as well as cooking aerosols (e.g., the effluent from hot cooking oil and bacon).

A small number of previous studies focused on cooktop fire sources and considered multi-detector sensing of pre-ignition signatures in a kitchen environment. Johnsson [6] conducted a series of experiments investigating the feasibility of distinguishing between normal cooking activities and preignition conditions using a variety of sensors in a mock kitchen with a closed door. Sensors were placed above the cooktop and on the compartment ceiling. Signals from

Mensch, Amy; Hamins, Anthony; Markell, Kathryn.
"Development of a Detection Algorithm for Kitchen Cooktop Ignition Prevention."
Paper presented at Suppression, Detection and Signaling Research and Applications Symposium (SUPDET 2018), Cary, NC, United States.
September 11, 2018 - September 14, 2018.

alcohol, CO, and hydrocarbon sensors showed potential to predict ignition while discriminating from normal cooking. Nearly all the experiments were conducted with the range hood off and the effects of room configuration and transport likely played a significant role in the interpretation of results. More recently, Johnsson and Zarzecki [7] conducted experiments which suggested that modified photoelectric smoke detectors could be used to warn of pre-ignition conditions while not impacting normal cooking scenarios.

Jain et al. [8] conducted cooking-oil autoignition experiments, considering the effectiveness of various inexpensive sensors to detect pre-ignition conditions, and reported that the rate-of-change of the moving average of CO concentration was a robust indicator of impending ignition. The study, however, did not consider normal cooking or common nuisance sources. The objective of our study was to determine which sensors/sensor combinations showed potential for use as input to a detection algorithm for cooktop ignition prevention. The initial set of experiments were focused on sensor response and were designed to limit transport considerations.

Experimental Apparatus and Procedures

In this study, ignition and normal cooking tests were conducted in a mock-up kitchen. Cooking oils were heated in a pan on an electric-coil stovetop with the highest power setting until ignition occurred. These tests used round, cast iron, aluminum, multi-layered, and stainless-steel pans with diameters of either 20 cm (8 in) or 25 cm (10 in). In most tests, the pans were placed in the rear locations on the cooktop, with the small burner used for the 20 cm pan and the large burner used for the 25 cm pan. On the highest setting, the stove power was about 1.1 kW on the small burner and 1.8 kW on the large burner.

Soybean, canola, olive, sunflower, and corn oils were tested, since these are commonly used cooking oils in the U.S [9]. Butter was also heated to ignition in one test. Normal cooking or nuisance sources included boiling water (steam), cooking hamburgers (80 % lean), and cooking seasoned salmon with butter. For the salmon cooking, the butter was heated on high for 3 min, the salmon was added and heated on high for 4 min, and the salmon was flipped and cooked on high for 4 min. Following that procedure, unattended cooking was simulated by continuing to cook the salmon at the high-power setting. In one case, the salmon eventually ignited. The cooking procedure for the hamburgers was the same as in Ref. [10]. Two hamburgers were also cooked in the oven on the broil setting according to the UL 217 Cooking Nuisance Smoke Test procedure [11]. A list of the experimental conditions is presented in Table 1.

Approximately 20 different sensor responses were selected for testing, including types that were based on various operating mechanisms, including electrochemical, catalytic, MOS-type, light scattering, and ionization. Sensors were selected to measure CO_2, CO, hydrocarbons, alcohols, H_2, natural gas, volatile organic compounds (VOCs), smoke, air quality, and aerosols/dust. Temperature and humidity were also measured. The dust sensor was modified twice to extend its range of sensitivity, and the dust-sensor iteration (1, 2 or 3) is listed in Table 1. The sensors were positioned approximately 3 m downstream of the exhaust duct inlet, which was located about 0.8 m above the cooktop. Data were acquired at ¼ Hz. The exhaust fan was set to high flow (about 3.4 m/s) in the duct. Part way into testing, aluminum foil was added to partially enclose the area from the cooktop up to the exhaust hood on the left and right sides. The partial enclosure ensured that most of the plume of hot aerosols and gases flowed into the hood and past the sensors stationed in the duct. In this way, it was possible to eliminate transport effects from consideration in interpretation of the experimental results after the aluminum foil was added, for experiments 8 – 15 and 18 – 33.

Mensch, Amy; Hamins, Anthony; Markell, Kathryn.
"Development of a Detection Algorithm for Kitchen Cooktop Ignition Prevention."
Paper presented at Suppression, Detection and Signaling Research and Applications Symposium (SUPDET 2018), Cary, NC, United States.
September 11, 2018 - September 14, 2018.

Table 1. List of Experimental Conditions

Experiment	Pan Type	Pan Diameter	Food and Amount	Burner Location	Burner Size	Foil	Dust Sensor
1, ignition	cast iron	20 cm	50 mL canola oil	rear	small	no	1
2, ignition	cast iron	20 cm	50 mL canola oil	rear	small	no	1
3, ignition	cast iron	20 cm	50 mL canola oil	rear	small	no	2
4, ignition	cast iron	20 cm	50 mL canola oil	rear	small	no	2
5, ignition	cast iron	20 cm	50 mL canola oil	rear	small	no	2
6, ignition	cast iron	20 cm	50 mL canola oil	rear	small	no	2
7, ignition	cast iron	20 cm	50 mL canola oil	rear	small	one side	2
8, ignition	cast iron	20 cm	50 mL canola oil	rear	small	yes	3
9, ignition	cast iron	20 cm	100 mL canola oil	rear	small	yes	3
10, ignition	aluminum	20 cm	50 mL canola oil	rear	small	yes	3
11, ignition	multi-layered	20 cm	50 mL canola oil	rear	small	yes	3
12, ignition	stainless steel	20 cm	50 mL canola oil	rear	small	Yes	3
13, ignition	cast iron	20 cm	200 mL canola oil	rear	small	yes	3
14, ignition	cast iron	20 cm	50 mL canola oil	rear	small	yes	3
15, ignition	cast iron	25 cm	100 mL canola oil	rear	large	yes	3
16, ignition	aluminum	20 cm	50 mL corn oil	rear	small	no	1
17, ignition	cast iron	20 cm	50 mL corn oil	front	small	no	2
18, ignition	cast iron	25 cm	100 mL corn oil	rear	large	yes	3
19, ignition	cast iron	20 cm	50 mL corn oil	rear	small	yes	3
20, ignition	cast iron	20 cm	50 mL soybean oil	rear	small	yes	3
21, ignition	cast iron	25 cm	100 mL soybean oil	rear	large	yes	3
22, ignition	cast iron	20 cm	50 mL olive oil	rear	Small	yes	3
23, ignition	cast iron	25 cm	100 mL olive oil	rear	large	yes	3
24, ignition	cast iron	25 cm	100 mL sunflower oil	rear	large	yes	3
25, ignition	cast iron	20 cm	50 mL sunflower oil	rear	small	yes	3
26, ignition	cast iron	20 cm	50 mL butter	rear	small	yes	3
27, normal cooking	broiler pan	N/A	2 x 230 g (0.5 lb) hamburgers	oven	N/A	yes	3
28, normal cooking	cast iron	20 cm	230 g (0.5 lb) hamburger	rear	small	yes	3
29, normal cooking	cast iron	25 cm	2 x 230 g (0.5 lb) hamburgers	rear	large	yes	3
30, normal cooking	cast iron	20 cm	227 g (8 oz) salmon, 47 mL butter	rear	small	yes	3
31, ignition	cast iron	20 cm	227 g (8 oz) salmon, 47 mL butter	rear	small	yes	3
32, normal cooking	cast iron	25 cm	454 g (16 oz) salmon, 93 mL butter	rear	large	yes	3
33, normal cooking	cast iron	20 cm	50 mL water	rear	small	yes	3

Mensch, Amy; Hamins, Anthony; Markell, Kathryn.
"Development of a Detection Algorithm for Kitchen Cooktop Ignition Prevention."
Paper presented at Suppression, Detection and Signaling Research and Applications Symposium (SUPDET 2018), Cary, NC, United States.
September 11, 2018 - September 14, 2018.

Results and Discussion

Figure 1 shows the transient pan temperatures during a typical experiment with oil ignition. For the oils, ignition occurred between 630 s and 880 s after the cooktop was powered, when the pan temperature was between 410 °C and 470 °C, consistent with previous studies [12]. As expected, the transient CO_2 signal was fairly flat until ignition, when it sharply increased as seen in Figure 2. Figure 3 shows many of the rest of the sensor signals during experiment 8, with each signal normalized by its own peak, which occurred near the time of ignition. Each sensor signal was characterized by a unique profile with its absolute value and slope varying in time. Several of the sensors appeared to provide signals that may be useful for providing early detection of impending ignition, including the sensors sensitive to dust, CO, and VOCs.

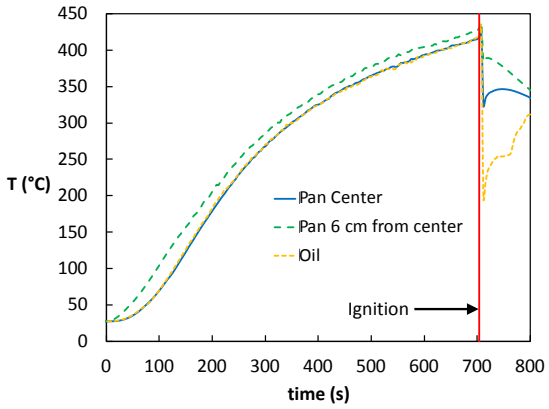


Figure 1. Pan temperatures for exp. 8.

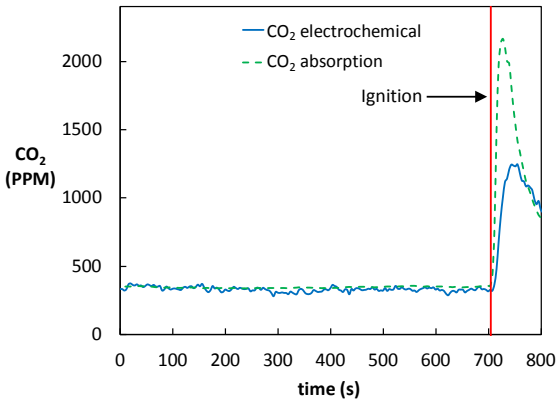


Figure 2. CO_2 measurements for exp. 8.

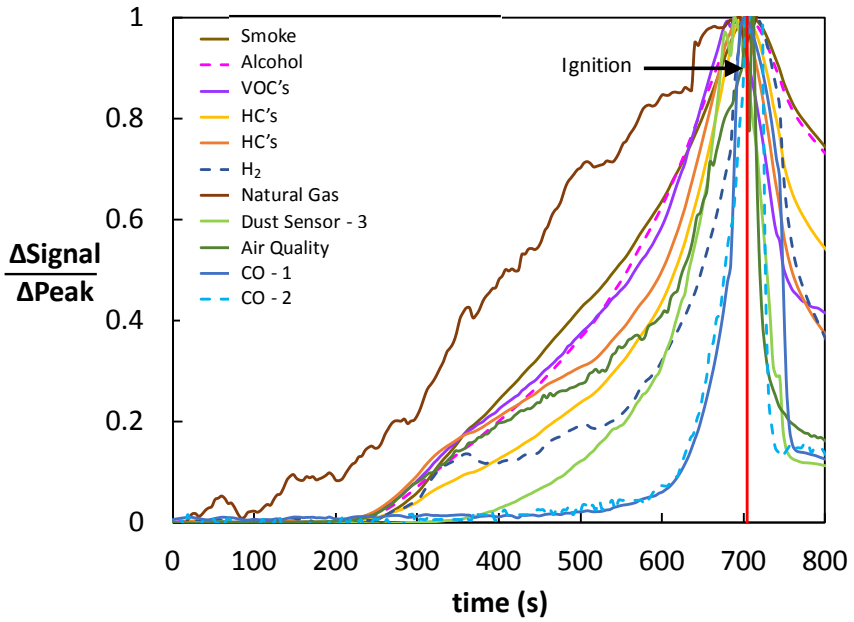


Figure 3. Normalized signals for exp. 8.

Mensch, Amy; Hamins, Anthony; Markell, Kathryn.
"Development of a Detection Algorithm for Kitchen Cooktop Ignition Prevention."
Paper presented at Suppression, Detection and Signaling Research and Applications Symposium (SUPDET 2018), Cary, NC, United States.
September 11, 2018 - September 14, 2018.

If a practical detection algorithm for cooktop ignition prevention is to be developed, detection must occur sufficiently before ignition to allow time to provide warning and/or direct action (e.g., cut the power), while also not impeding normal cooking activities. A time of 60 s before ignition is judged as a minimum time for a detection algorithm target, partly due to thermal inertia of the pan-cooktop system [6]. Many of the sensors, including the dust sensor and VOCs sensor, output a voltage reading, which has not been calibrated to concentration or other measurements at this point. If the value of one of these sensors is used in a proposed algorithm, these signals could be calibrated to be able to directly compare to other sensors.

Figure 4 compares the dust-sensor signal (minus the background signal), in volts, for all the experiments with a focus on two moments in time, namely 60 s before ignition and at the time of the peak signal. The results for the ignition of oils and butter are shown on the left portion of the graph. Normal cooking cases and unattended salmon ignition are on the right, with normal cooking results outlined by the box at the bottom right. A dust-sensor signal output of about 0.5 V seemed to distinguish the oil results from the normal cooking results except for one normal cooking experiment and one cooking-oil experiment (highlighted in the figure with unfilled squares).

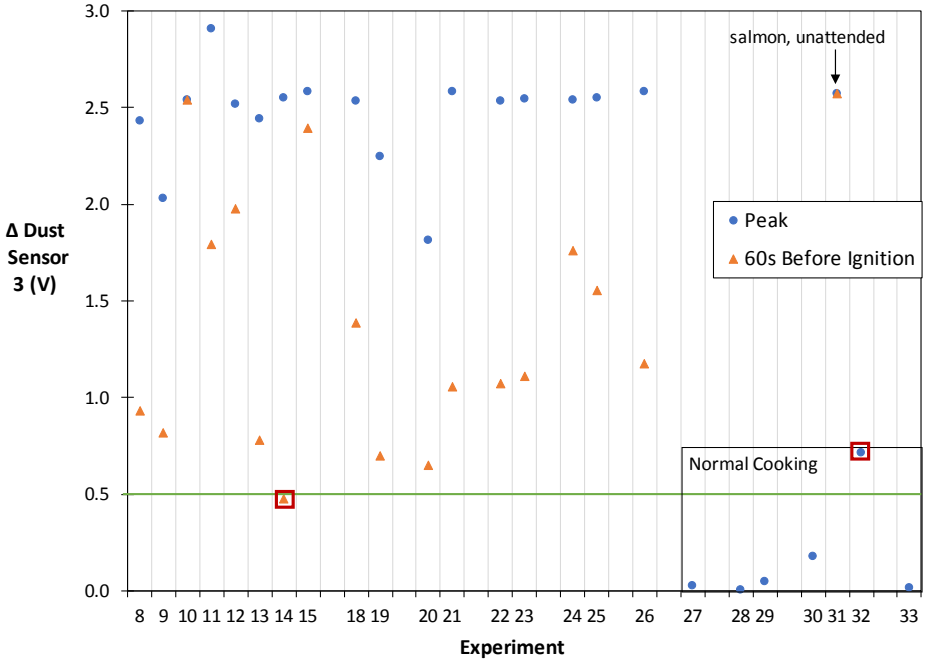


Figure 4. Change in dust sensor signal for experiments using configuration 3: peak and at 60 s before ignition.

Figure 5 shows analogous results for the time rate-of-change of CO. The figure compares the raw values of d(CO)/dt for all the experiments with a focus on 60 s before ignition and at the time of the peak signal. The results for the oils are shown on the left portion of the graph; normal cooking results are shown on the right. Since the CO signal tends to increase very rapidly close to ignition, the derivative values are plotted on a log axis. A derivative value on the order of d(CO)/dt = 0.6 generally seemed to distinguish the ignition results from the normal cooking results. This value would not prevent any of the normal cooking activities, but it would not catch two tests with

oil ignition until < 60 s before ignition. For experiments 5 and 6, d(CO)/dt would reach 0.6 V 35 s before ignition and at ignition, respectively.

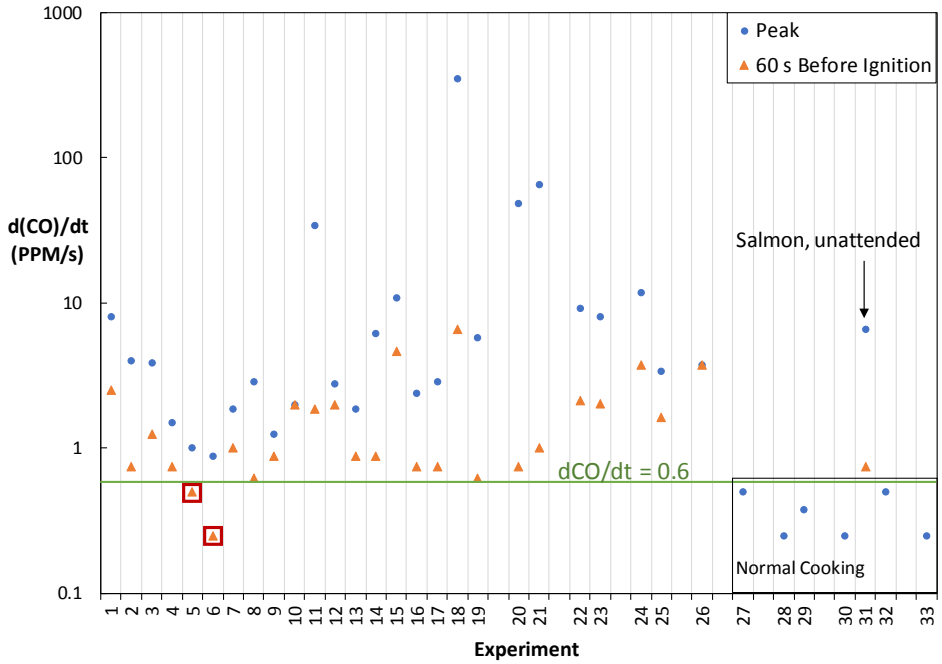


Figure 5. Rate of change of CO: peak and at 60 s before ignition.

Figure 6 shows analogous results for the change in the VOC signal from the background, in volts, comparing the peak signal and 60 s before ignition for all the experiments. A value on the order of 0.55 V appeared to differentiate the results for normal cooking from the cases with ignition. This criterion would have no false positives within 60 s of ignition and no false negatives with normal cooking for the experiments in this test series. However, there is not a large difference between the cutoff point and the highest signals from the normal cooking cases. Without additional repeat experiments of these cases and other similar experiments to determine the variability of these signals, we cannot be sure of the robustness of this algorithm to prevent all false alarms and ignitions.

Summary and Conclusions

A series of experiments was conducted to investigate the possibility of sufficiently early detection of imminent ignition during cooking. The results suggest that 1) a variety of sensors are sensitive to the plume of gases and aerosol associated with cooking. and 2) a number of algorithms show promise in distinguishing imminent (within 60 s) ignition from normal cooking activities, particularly with sensors that detect dust, CO and VOC's. Further work is needed to determine the variability of the sensor signals under a broader set of realistic conditions that encompasses sensor location and hood fan flow. This would test the robustness of current algorithms, as well as other algorithms incorporating other signals or additional signals. It would be beneficial to consider algorithms that are transport independent. Possible transport independent algorithms could be the time rate-of-change of a sensor, a ratio of the signals from two different sensors, or the derivative

Mensch, Amy; Hamins, Anthony; Markell, Kathryn.
"Development of a Detection Algorithm for Kitchen Cooktop Ignition Prevention."
Paper presented at Suppression, Detection and Signaling Research and Applications Symposium (SUPDET 2018), Cary, NC, United States.
September 11, 2018 - September 14, 2018.

of one sensor signal with respect to another sensor signal. Additionally, tests will need to be conducted to ensure that if the stove power is shut off when the condition(s) of a certain algorithm have been met, that ignition is prevented.

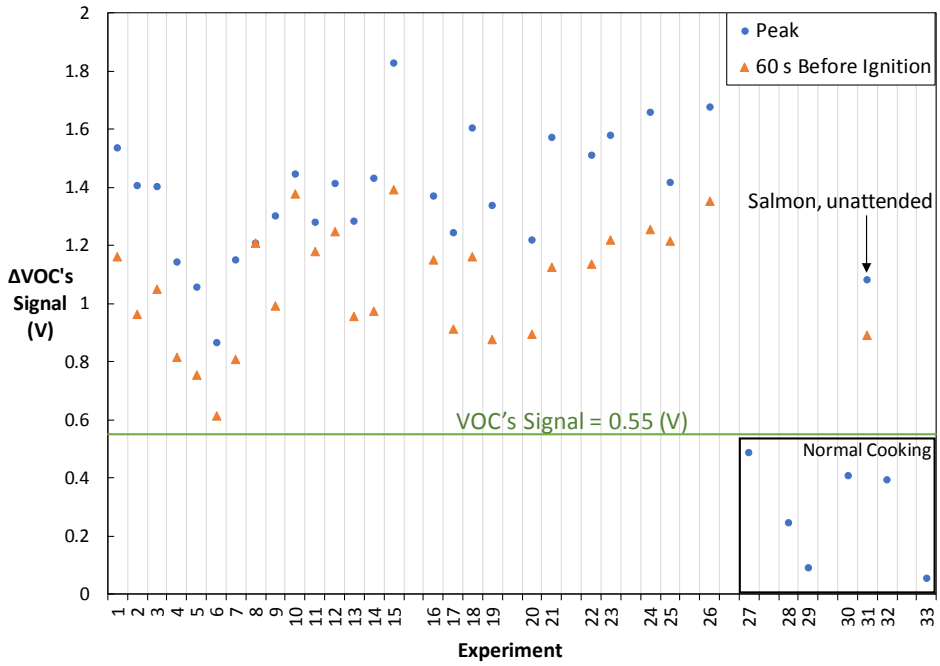


Figure 6. Change in VOC signal: peak and at 60 s before ignition.

References

[1] M. Ahrens, Home Fires Involving Cooking Equipment, National Fire Protection Association, Quincy, MA(2017).

[2] Underwriter's Laboratory, Northbrook, IL, Standard for Household Electric Ranges, Underwriter's Laboratory, Northbrook IL **UL 858** (2014).

[3] D.T. Gottuk, M.J. Peatross, R.J. Roby, and C.L. Beyler, Advanced fire detection using multi-signature alarm algorithms, Fire Safety Journal **37** (4), 381–394 (2002).

[4] L.A. Cestari, C. Worrell, and J.A. Milke, Advanced fire detection algorithms using data from the home smoke detector project, Fire Safety Journal **40** (1), 1–28 (2005).

[5] D.T. Gottuk, M.T. Wright, J.T. Wong, H.V. Pham, S.L. Rose-Pehrson, S. Hart, M. Hammond, F.W. Williams, P.A. Tatem, and T.T. Street, Prototype Early Warning Fire Detection System: Test Series 4 Results, (2002).

[6] E.L. Johnsson, Study of Technology for Detecting Pre-Ignition Conditions of Cooking-Related Fires Associated with Electric and Gas Ranges and Cooktops, Final Report, National Institute of Standards & Technology, Gaithersburg, MD(1998).

[7] E. Johnsson and M. Zarzecki, Using Smoke Obscuration to Warn of Pre-Ignition Conditions of Unattended Cooking Fires, *16th International Conference on Automatic Fire Detection (AUBE '17) & Suppression, Detection and Signaling Research and Applications Conference (SUPDET 2017)*, (2017).

[8] A. Jain, P. Nyati, N. Nuwal, A. Ansari, C. Ghoroi, and P. Ghandi, Pre-Detection of Kitchen Fires due to Auto-Ignition of Cooking Oil and LPG Leakage in Indian Kitchens, Fire Safety Science **11** 1285–1297 (2014).

[9] M. Ash, Edible fats and oils: U.S. Supply and disappearance, 2002/03-2015/16, Economic Research Service, United States Department of Agriculture, (2016).

[10] T.G. Cleary, A study on the performance of current smoke alarms to the new fire and nuisance tests prescribed in ANSI/UL 217-2015, National Institute of Standards and Technology, Gaithersburg, MD(2016).

Mensch, Amy; Hamins, Anthony; Markell, Kathryn.
"Development of a Detection Algorithm for Kitchen Cooktop Ignition Prevention."
Paper presented at Suppression, Detection and Signaling Research and Applications Symposium (SUPDET 2018), Cary, NC, United States.
September 11, 2018 - September 14, 2018.

[11] Underwriter's Laboratory, Northbrook, IL, Standard for Safety Smoke Alarms, **ANSI/UL 217** (2015).
[12] J.B. Dinaburg and D.T. Gottuk, Development of Standardized Cooking Fires for Evaluation of Prevention Technologies: Data Analysis, National Institute of Standards and Technology, (2015).

Mensch, Amy; Hamins, Anthony; Markell, Kathryn.
"Development of a Detection Algorithm for Kitchen Cooktop Ignition Prevention."
Paper presented at Suppression, Detection and Signaling Research and Applications Symposium (SUPDET 2018), Cary, NC, United States.
September 11, 2018 - September 14, 2018.

26th CIRP Life Cycle Engineering (LCE) Conference

Incorporating unit manufacturing process models into life cycle assessment workflows

William Z. Bernstein[a,*], Cesar D. Tamayo[b], David Lechevalier[c], Michael P. Brundage[a]

[a]*Systems Integration Division, NIST, 100 Bureau Dr, Gaithersburg, MD 20899, USA*
[b]*Ira A. Fulton School of Engineering, Arizona State University, 699 S Mill Ave, Tempe, AZ 85281, USA*
[c]*Engisis LLC, 10411 Motor City Dr Ste 750, Bethesda, MD 20817, USA*

Abstract

Life cycle assessment (LCA) carries significant uncertainties and imprecision due to a number of factors, including the framework's linearity assumptions and the wide use of aggregate unit processes in practice. In this work, we exploit the unit manufacturing process (UMP) information model (ASTM E3012-16) to enable parametric environmental analysis of manufacturing systems without disrupting the traditional LCA workflow. We present a formal mapping of an extension of the ASTM E3012 data model and the ecoSpold2 data model. We then demonstrate the utility of this mapping by (1) generating life cycle inventory (LCI) data from an example UMP model representing a vertical milling process and (2) linking the results with an existing LCI database. To show value, we use the Brightway2 framework to process the LCI data and complete a LCA. We conclude by comparing LCA results generated from the parametric milling UMP model against LCA results of a similar milling unit process model from a commercial database.

Peer-review under responsibility of the scientific committee of the 26th CIRP Life Cycle Engineering (LCE) Conference.

Keywords: Unit Manufacturing Process, Life Cycle Inventory Data, Life Cycle Assessment, Brightway2, Jupyter Notebooks, Ecoinvent, ASTM E3012

1. Introduction

Current life cycle assessment (LCA) practices carry significant uncertainty due to a lack of data and reusable parametric models as well as the presence of a number of critically flawed assumptions (e.g., models are linear to single inputs and are transferable across similar geographical locations) [20]. Life cycle inventory (LCI) database (*"pre-computed"*) models contain methods, e.g., the pedigree matrix, to deal with such uncertainties, yet the debate on their efficacy continues (see recent editorial from Heijungs et *al*. [10]). As a result, practitioners seeking more precise, scalable, and parametric LCI models spend significant effort in constructing their own models from scratch [7]. Without a standard model representation, it has become increasingly more difficult to properly exchange, reproduce, and explain LCA workflows. Leading researchers and practitioners have recognized these challenges through a recent LCA capability roadmap, stating that three of the most critical opportunities are *describing model contents*, *describing model structure*, and *collaborative use of models* [15]. In the LCA community, some have attempted to improve the transparency of their work by including supplemental material describing their models [5, 21]. However, manual reconstruction is still required to replicate these studies.

In response to these challenges, this paper leverages an existing standard for representing parametric manufacturing process models, i.e., unit manufacturing process (UMP) models as defined by ASTM E3012-16 [1], and links them to traditional LCA workflows. By generating LCI data from UMP models, we demonstrate a means for storing and exchanging parametric LCI models for manufacturing processes. Manufacturing processes present a key opportunity since existing manufacturing LCI models available in commercial databases do not commonly feature process-level parametric relationships to enable decision making in traditional manufacturing workflows. Instead, models rely on high-level aggregated assumptions that are not scalable to low-level manufacturing operations, e.g., distinguishing differences in milling slots or pockets of the same

* Corresponding author. Tel.: +1-301-975-3528 ; fax: +1-301-975-9749.
E-mail address: wzb@nist.gov (William Z. Bernstein).

Bernstein, William; Tamayo, Cesar; Lechevalier, David; Brundage, Michael.
"Incorporating unit manufacturing process models into life cycle assessment workflows."
Paper presented at 26th CIRP Conference on Life Cycle Engineering (LCE), West Lafayette, IN, United States. May 7, 2019 - May 9, 2019.

volume. For example, the ecoinvent database[1] scales all machining operations based on the weight of product produced or operation conducted. In other words, removing material in a complex manner to create a 1 kg sphere and cutting simpler shapes to create a 1 kg cube would share identical environmental impacts. Such an assumption is fundamentally flawed and causes significant uncertainty.

Motivated by such challenges, others have developed frameworks and tools to develop, curate, and deploy parametric process models to achieve more sustainable manufacturing [8, 9, 11, 13, 14, 19]. However, these solutions do not follow a strict standards-based approach and are hence difficult to integrate into traditional LCA workflows. Our approach is complementary to these efforts yet maintains a strong focus on standards throughout its design and implementation.

Both E3012-16 and ISO14048 [12] contain data representations for representing environmental impacts of manufacturing processes[2]. However, these representations are incompatible due to their differing purposes. E3012-16 is designed to communicate and formally characterize the performance of manufacturing processes through a common information model while the ecoSpold2 format was created to curate LCI datasets for LCAs in databases and conforms to ISO 14048, which defines requirements for LCA data formats [12]. Providing a model transformation from E3012 into LCA workflows allows for more accurate manufacturing process models to be considered when conducting an LCA. This would allow manufacturers to reuse their production models in LCAs and would allow LCA practitioners to better understand how a change at the production phase could ripple throughout the entire product lifecycle. This paper explores this model transformation by mapping an E3012 model[3] into ecoSpold2 and conducting an LCA using the Brightway2 framework [17]. Note that the E3012 model encodes *UseBounds* for each model input and output, facilitating record-keeping related to uncertainty quantification [2].

In this paper, we present (1) the development of a formal mapping between the UMP and ecoSpold2 information models, (2) the generation of LCI data demonstrated through a milling case study, and (3) guidance for the revision of E3012 to facilitate its utility in LCA workflows. We view this work as critical in (a) forming a bridge between previous efforts of curating parametric manufacturing models, such as the Cooperative Effort on Process Emissions in Manufacturing (CO2PE!) [7] and (b) presenting a cohesive vision for a UMP repository [3].

2. Methods and tools deployed

The main goal of this work is to develop a pipeline that ports data from the UMP representation into the traditional LCA workflow. Here, we assume that users are implementing the revised E3012 information model [2] to communicate and exchange UMP models. We also assume that LCA software in our workflow accepts the ecoSpold2 information models as inputs. Based on prior experience, we also assume that user input is necessary due to the required domain expertise of selecting LCI models from traditional databases. Requirements for achieving the mapping between the UMP and ecoSpold2 formats in an open-source manner, include (1) an open tool that accepts and runs UMP models, (2) open LCA software that ports to LCI models, and (3) an interactive framework that prompts practitioners for domain expertise when needed.

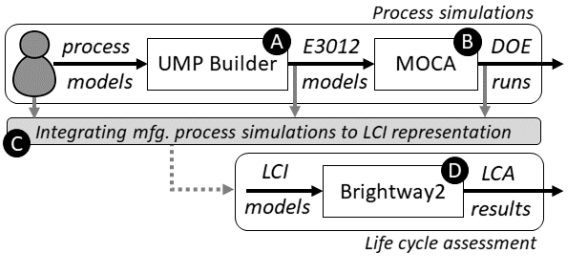


Fig. 1. Pipeline realized by mapping E3012 to ecoSpold2. Each labeled step (A-D) signifies a stage of data transformation or manipulation (A-D).

Figure 1 presents the data pipeline for generating LCA results from parametric models curated as UMP models. To begin, a manufacturer or modeler contributes a parametric model representing a manufacturing process. In our work, ASTM E3012 is used to represent domain-specific data about the physical inputs, outputs, and resources, as well as mathematically defined transformations and product and process information [1]. We leverage the UMP Builder [2] (labeled as A in Fig. 1) to help manufacturers validate their conformance to the standard, share and reuse their UMP models, as well as interface with modeling, simulation, and analysis tools.

From the UMP model, we extract the structure and content to obtain operational code by using the **MO**del **C**omposition and **A**nalysis (MOCA) tool [16] (B in Fig. 1), outputting a Jupyter notebook[4]. This code contains control parameters set by the manufacturer and variable constraints that enable bounded simulations. The output code from MOCA can also be used for optimization, which could help improve a system with respect to a given metric of interest, e.g., cost or energy consumption. Executing the simulation generates a text file that stores all the values involved in each of the instances, e.g., control parameters, intermediate variables and metrics of interest.

Using both the parametric model and the simulation results, we perform a user-assisted mapping (C in Fig. 1) that yields an ecoSpold2 file compatible with Brightway2 (D in Fig. 1), an LCA framework. This file contains not only data describing the physical input and output of the manufacturing process in question but also links to other entries that provide inventory data of processes involved. This improves precision of results by covering the complete life cycle of the product. The generated ecoSpold2 file is then added to a dataset to be used by

[1] We considered ecoinvent 3.4 (see: https://www.ecoinvent.org/)
[2] The LCA data format used in this paper is ecoSpold2
[3] We use the schema extension proposed in Bernstein et *al.*[2]. The extension has been proposed as a E3012 revision and is under ballot in ASTM E60.13.
[4] Used for evaluating the UMP models (see: http://jupyter.org/)

Bernstein, William; Tamayo, Cesar; Lechevalier, David; Brundage, Michael.
"Incorporating unit manufacturing process models into life cycle assessment workflows."
Paper presented at 26th CIRP Conference on Life Cycle Engineering (LCE), West Lafayette, IN, United States. May 7, 2019 - May 9, 2019.

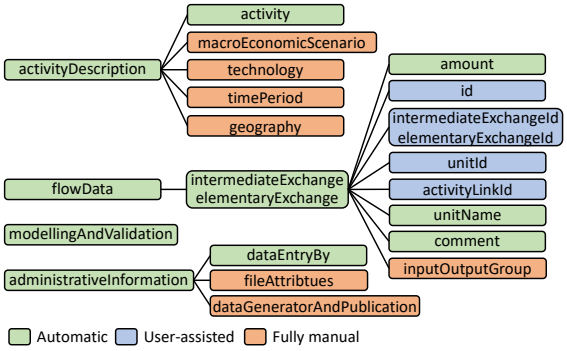


Fig. 2. Necessary fields for a generated ecoSpold file. Color of each field classifies how information is ported from UMP models in our implementation.

Brightway2 for performing LCA, assuming that the ecoSpold2 file has been appropriately generated.

To be clear, Fig. 1-[A, B, and D] represent steps that are generally applicable to other scenarios. The UMP Builder [2], can be used to generate models conforming to the revised E3012 schema. MOCA [16] can be used to graphically develop operational models using a domain-specific modeling language and Brightway2 [17] can be used to conduct LCAs. Our mapping (Fig. 1-C) facilitates the correlation between all three tools.

3. Mapping between the E3012 and ecoSpold2 data formats

Even though there are similarities between the E3012 and ecoSpold2 formats, we identified major differences that involved necessary steps for validation to successfully append the LCI database, e.g., *exchangeIds* and *unitIds*. Figure 2 classifies the mandatory fields for an ecoSpold2 file to be accepted by Brightway2 based on whether the information is available from the UMP model or additional support is required.

With the data provided by the UMP model and the simulations from MOCA, some of the required fields to generate valid ecoSpold2 files can be directly populated (Fig. 2, in green). However, in other cases, the system prompts the user to select an equivalent entry in the database (Fig. 2, in blue). For example, if the UMP model includes aluminum scrap as an output, the user must specify the appropriate option available in the LCI database, e.g., "treatment of aluminium scrap, post-consumer, prepared for recycling, at remelter" or "treatment of aluminium scrap, new, at remelter" as in ecoinvent. Linking inappropriate activities can significantly impact the LCA results. Data types shown in orange in Fig. 2 signify ecoSpold2-specific information not currently represented in E3012:

- *technology*, capturing characterizations of the technological domain of the activity, e.g., the relative modernity and significant peculiarities of the domain
- *macroEconomicScenario*, allowing for alternative macro-economic activities to be modelled and captured
- *dataGeneratorAndPublication*, containing information about who collected, compiled, or published the data, which may be the same person as under *dataEntryBy*
- *inputOutputGroup*, providing more details by classifying them into categories, such as materials/fuels, electricity/heat, services, or activities from the technosphere

For these instances, we added dummy data to meet ecoSpold2 requirements. These additions do not affect the LCA results.

To accomplish the mapping, the Mapping Module (MM) first extracts the input and output names and symbols (captured as MathML equations) from the E3012 model. For each input and output, a corresponding exchange in ecoSpold2 will be created. The MM uses the symbols to extract the input and output quantities computed by the MOCA simulation and maps them to the amount of the respective exchange.

To perform a LCA, each exchange needs to be linked to an activity. Inputs of the UMP model are linked to exchanges that are produced by an activity while outputs of the UMP model are linked to exchanges consumed by an activity. The exchange representing the reference product, i.e., the resulting product of the manufacturing process, does not need to be linked. Since the database could contain thousands of datasets, identifying the appropriate activity to link is user-assisted. For an exchange used as an input, the MM will provide activities that contain a "reference product" exchange matching with the name of the E3012 input. For an exchange used as an output (excluding the reference product), the MM is going to provide activities that consume the exchange as input from the technosphere, and match with the name of the E3012 output. The user must choose the appropriate activity from the prompted list.

Since E3012 does not currently handle a way to specify a reference product, the user is prompted to specify which output relates to the product generated. The reference product exchange is treated differently since it needs not to be linked to an existing activity. Once the appropriate activity has been chosen, the MM generates a *IntermediateExchangeId* and includes a *activityLinkId*, which represents the *id* of the entry to be linked.

For the rest of the fields (Fig. 2, in orange), the user manually adds information during mapping. For example, the *geography* field can be instantiated by finding the appropriate geographical location in the meta-data files, e.g., the *id* corresponding to United States. A similar approach can be used for *timePeriod*, *macroEconomicScenario*, *dataGeneratorAndPublication*, and *fileAttributes*. Some fields such as *technology* and *modelingAndValidation* are required in the ecoSpold2 schema. However, Brightway2 does not use their content. In other words, dummy values added for these fields do not affect LCA results.

In our implementation, we assume that all physical inputs and outputs of manufacturing processes are received from or generated to the technosphere, representing activities generated by human-driven economic processes. In future work, we plan to enable mapping to activities and flows to and from the ecosphere, representing flows that directly interface with ecological systems (e.g., waste water into a river from a coal burning plant). However, this requires more detailed and structured information about the inputs and outputs in the E3012 data model.

Bernstein, William; Tamayo, Cesar; Lechevalier, David; Brundage, Michael.
"Incorporating unit manufacturing process models into life cycle assessment workflows."
Paper presented at 26th CIRP Conference on Life Cycle Engineering (LCE), West Lafayette, IN, United States. May 7, 2019 - May 9, 2019.

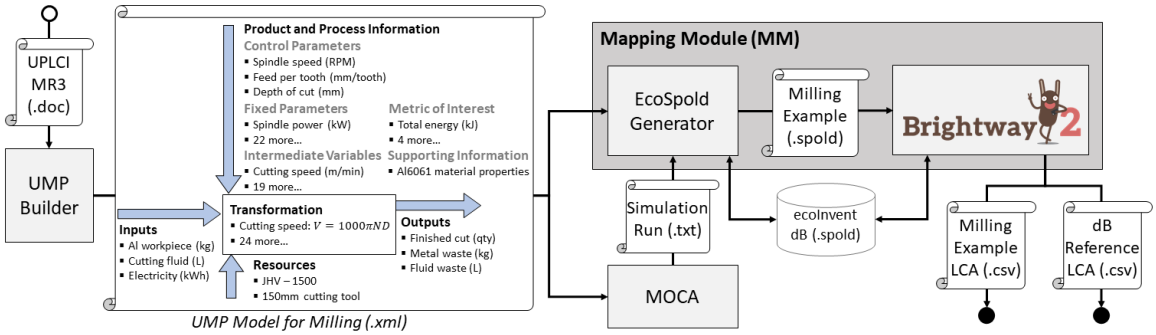


Fig. 3. Demonstration of the UMP-ecoSpold2 mapping through a milling case study.

4. Case study: integrating a milling UMP with ecoinvent

Figure 3 describes the data used and generated to demonstrate our UMP-ecoSpold2 mapping methodology. We borrow all assumptions and modeling procedures, including functional unit, scope, and system boundaries, from the milling example (code: MR3) reported by the Unit Process Life Cycle Inventory (UPLCI) team [18]. We built the model on the UMP Builder[5] and consulted the MR3 document as needed. Through the UMP Builder, an eXtensible Markup Language (XML) document was generated formally describing the parametric milling UMP model. This model consists of 25 transformation equations, 3 physical inputs, 3 physical outputs, 2 elements describing the manufacturing resources referenced in MR3, and a total of 52 entities describing product and process information. For every variable used in the transformation equations, an accompanying definition of its type, bound, and unit are captured under product and process information.

The semantic information describing each variable, equation, and the relationships between them is interpreted with the MOCA tool to generate operational code in the form of a Jupyter notebook. We used the MOCA-generated code to evaluate the UMP milling model. This case study presented the energy, waste, and time consumed for milling a straight cut of 90 000 mm^3 of prismatic aluminum (Al) workpiece. For the case study, the control parameters, depth of cut, spindle speed, and feed per tooth, were set to 3 mm, 255 rev/min, and 0.381 mm/tooth, respectively. We recognize that these settings are conservative; however, we aimed to conform exactly to the UPLCI model. All computed values were compared to the case study section of the UPLCI MR3 document to validate our milling model was created and evaluated appropriately.

To generate an ecoSpold2 file corresponding to the UMP model, the MM extracts semantic information from the milling XML document, including units, symbols, and names. The MM also obtains numerical data from the text file generated from MOCA, including values associated with metrics of interest (e.g., waste generated). Here, each representing metrics of interest, the computed energy consumption, cycle time, aluminum waste generated, and CO_2 emissions are 0.334 kWh, 90.3 s, 0.244 kg, and 0.196 kg CO_2, respectively. Through the use of the semantic data, the MM queries onto the ecoinvent database, a commercial LCI database, to link entities of the milling UMP, e.g., aluminum 6061, cutting fluid, and electricity, to database activities that either generates the UMP input or consumes the UMP outputs. This is necessary to perform a complete LCA. In the generated "MillingExample.spold" file, the functional unit of a single cut is set to a dimension of 90,000 mm^3 (or 0.24489 kg of Al). Energy consumed, waste generated, and cycle time were scaled based on the size of the cut. We rely on Brightway2 to evaluate the milling process's environmental impacts using the Tool for Reduction and Assessment of Chemicals and other Environmental Impacts (TRACI) methods[6]. After verifying that the milling UMP can be used to generate LCA data, we conducted an initial validation study to test whether our model is producing realistic, feasible values as compared with commercial models present in the ecoinvent 3.4 database. In this use case, we compare the values generated by our milling UMP against *aluminum milling, small parts RoW*, which is an entry in the ecoinvent database, since the metadata description within the ecoinvent file seemed to match the intent of the UPLCI MR3 milling descriptions. While comparing to the dataset *aluminum milling, small parts*, there were some important considerations. The ecoinvent database selects the weight of the material cut from the part as a functional unit, making the initial shape of the part a fundamental consideration in the equation. In our test case, we use a single horizontal cut instead, allowing us to obtain a more precise and scalable measure.

To compare results between the TRACI impacts of the milling case study with the existing database (DB) entry, we conducted nine Monte Carlo (MC) simulations (50 000 runs each) with the Brightway2 framework using the uncertainty properties from the ecoSpold2 file of the DB entry. The main idea was to perturb each individual exchange of the *aluminum, small parts RoW* based on their individual uncertainty characteristics, fit a probability density function (PDF) to the results

[5] Public version of UMP Builder (see: https://umpbuilder.nist.gov/)

[6] TRACI was developed by the Environmental Protection Agency (EPA). See https://tinyurl.com/yde3bjno

Bernstein, William; Tamayo, Cesar; Lechevalier, David; Brundage, Michael.
"Incorporating unit manufacturing process models into life cycle assessment workflows."
Paper presented at 26th CIRP Conference on Life Cycle Engineering (LCE), West Lafayette, IN, United States. May 7, 2019 - May 9, 2019.

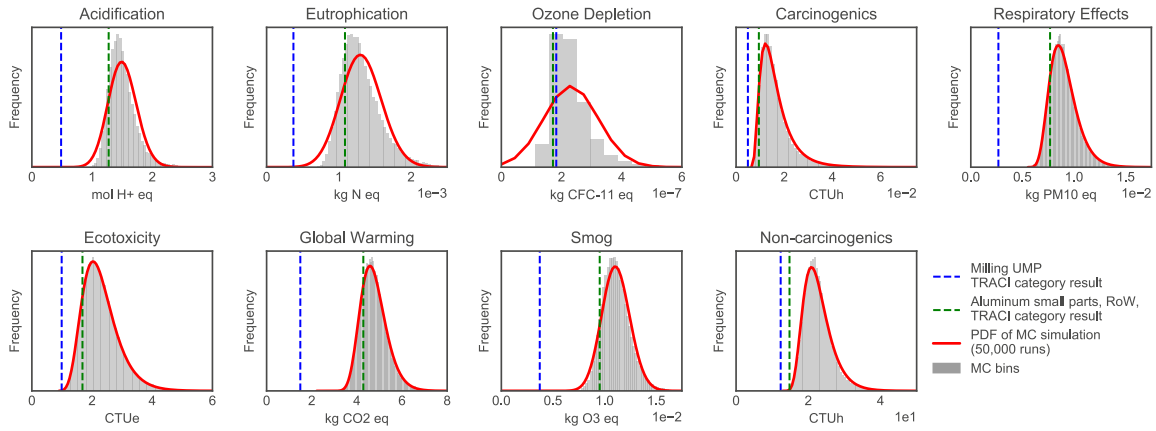


Fig. 4. Results of a Monte Carlo simulation (50,000 runs) of *aluminum, small parts RoW* with comparison of results to our Milling UMP (blue dotted line). The green dotted line signifies results of an LCA conducted with only the nominal values available in the DB entry.

Table 1. Our test case compared against similar activity in ecoinvent
*Refers to values from database entry, *aluminum milling, small parts RoW*

TRACI category (units)	DB*	CDF_{DB}*	UMP	CDF_{UMP}
acidification (mol H+ eq)	1.28	0.162	0.486	6.32e-7
ecotoxicity (CTUe)	1.68	0.132	0.990	8.53e-5
eutrophication (kg N eq)	1.10e-3	0.213	3.71e-4	2.22e-4
global warming (kg CO_2 eq)	4.29	0.214	1.49	**0.0**
ozone dep. (kg CFC-11 eq)	1.72e-7	0.251	1.83e-7	0.292
smog (kg O_3 eq)	9.53e-3	0.121	3.72e-3	5.41e-10
carcinogenics (CTUh)	9.47e-3	6.30e-2	4.91e-3	**0.0**
non-carcinogenics (CTUh)	14.8	3.39e-4	12.4	3.15e-14
resp. effects (kg PM10 eq)	7.68e-3	0.143	2.67e-3	**0.0**

based on the TRACI categories, and observe if our test case data falls within the bounds of the PDF. According to the DB entry, each exchange is modeled as a lognormal random variable. Here, we assume that the MC results can be approximated as a lognormal distribution. Though difficult to prove, it has been observed that linear combinations of lognomal random variables effectively approximate to a lognormal distribution [6].

Figure 4 summarizes the result of the nine MC simulation runs for each TRACI impact category. The PDFs fitted to the simulation data, the values of the milling UMP test case, and the values of the DB entry are shown in red, blue, and green, respectively. The values from the UMP results are considerably lower than those for the database entry, except for results for ozone depletion. To understand the degree of their difference, we evaluated the cumulative distribution function (CDF) at each value, as shown in Table 1. As seen in the CDF evaluation for the UMP values, with the exception of the ozone depletion result, the UMP results fall outside the uncertainty bounds of the database entry. In other words, the CDF evaluations are practically zero. In three cases, i.e., global warming, carcinogenics, and respiratory effects, the evaluation of the CDF was zero (shown in bold). Interestingly, the discrete values from the database entry seem to represent a rather liberal estimation of the results, falling to the left tail of the PDF.

Here, we offer an explanation for the differences observed. The complexity of both models are considerably different. The UMP milling example carries 6 exchanges while the database entry has 27 exchanges. If we were to include, for example, impacts associated with compressed air and other auxiliary manufacturing resources (similar to the ecoinvent entry), we would expect to obtain closer values. However, it is not clear which of the 9 resulting values (i.e., which impact category) would be most affected. These issues get to the center of the difference between a parametric approach and using "pre-computed" LCI data. The "pre-computed" data is heavily aggregated and incorporates effects from industry-wide exchanges regardless of whether the process utilizes every one. This is evident in the low CDF values of the database entry itself against the MC simulations. However, we recognize that parametric models built using the E3012 data model require more rigorous testing and validation than what was done for the milling UMP example. Characterizing the validation requirements of such UMP models to be as trusted as "pre-computed" LCI models is a necessary step to push this work forward.

5. Future directions and closing remarks

In this paper, we discussed the mapping of the E3012 and ecoSpold2 data models and demonstrated its utility in a traditional LCA workflow using Brightway2. Through this exercise, we informed the on-going revision of the E3012 standard. For example, we included units and bound equations for *Input* and *Output* entities in the UMP to ease the integration with LCA tools. We also identified an opportunity to integrate a definitive "functional unit" and clearer classifications of waste into the UMP information model. However, these concepts require additional research to be addressed properly.

Bernstein, William; Tamayo, Cesar; Lechevalier, David; Brundage, Michael.
"Incorporating unit manufacturing process models into life cycle assessment workflows."
Paper presented at 26th CIRP Conference on Life Cycle Engineering (LCE), West Lafayette, IN, United States. May 7, 2019 - May 9, 2019.

Our work is not without its limitations. Our pipeline relies on significant human input for some of the mapping, as discussed in Section 3. Selecting appropriate database entries is an expert-driven exercise and, hence, is prone to human error. Another limitation is that we do not yet integrate the design of experiments simulations from MOCA with Brightway2. In other words, we do not fully leverage the rich information describing the control variables to simulate LCA data. If such integration was realized, relating LCA results to product design decisions would be feasible. Additionally, we assume in this work that a single UMP model maps in a one-to-one fashion to a single LCI database entry. We do not address pooling information from multiple UMP sources to a single LCI process.

Other limitations of this work relate to the E3012 information model and support around it. As of now, we have yet to demonstrate validation protocols for UMP models. To integrate information from several UMP models, consistency in model topography is critical, including considerations related to naming conventions, units, and shared content (e.g., equations). Developing a "master data" context similar to how ecoinvent handles this issue could be a reasonable research direction.

To conclude, we plan to relate the LCI data generation back to the control variables defined in the UMP to enable system-tradespace exploration. One of the key challenges with effectively making environmentally-efficient decisions at the design stage is having the appropriate data representations speak to one another. From that perspective, previous design tools and frameworks have not been ideal [4]. We envision that integrating the UMP information model will help realize a new suite of tools that can explore "what-if scenarios" tied to design decisions and how their effects propagate through the lifecycle. In other words, we will extend the pipeline to relate UMP models to parametric design attributes. For example, how does the number of teeth in a gear design change the machining strategy and what is its impact on the environment? Developing a automated pipeline to reflect on such questions would facilitate deeper design space exploration. We believe that such an achievement would demonstrate the impact and scalability of the UMP modeling approach.

Disclaimer

No endorsement of any commercial product by NIST is intended. Commercial materials are identified in this report to facilitate better understanding. Such identification does not imply endorsement by NIST nor does it imply the materials identified are necessarily the best available for the purpose.

Acknowledgements

We thank Moneer Helu, Tesfaye Moges, and Chris Mutel for their valuable feedback that improved the paper. We also acknowledge Prof. Gabor Karsai, Amogh Kulkarni, and the ISIS Lab at Vanderbilt University, for implementing a UMP model parser easing its integration with MOCA, a WebGME tool.

References

[1] ASTM E3012-16, 2016. Standard Guide for Characterizing Environmental Aspects of Manufacturing Processes. ASTM International.
[2] Bernstein, W.Z., Lechevalier, D., Libes, D., 2018a. UMP Builder: Capturing and exchanging manufacturing models for sustainability, in: ASME 2018 International MSEC collocated with the 46th NAMRC, ASME.
[3] Bernstein, W.Z., et al., 2018b. Research directions for an open unit manufacturing process repository: A collaborative vision. Manufacturing Letters 15, 71–75.
[4] Brundage, M.P., Bernstein, W.Z., Hoffenson, S., Chang, Q., Nishi, H., Kliks, T., Morris, K., 2018. Analyzing environmental sustainability methods for use earlier in the product lifecycle. J CLEAN PROD 187, 877–892.
[5] Cheung, C.W., Berger, M., Finkbeiner, M., 2018. Comparative life cycle assessment of re-use and replacement for video projectors. The International Journal of Life Cycle Assessment 23, 82–94.
[6] Di Renzo, M., Graziosi, F., Santucci, F., 2009. Approximating the linear combination of log-normal rvs via pearson type iv distribution for uwb performance analysis. IEEE T COMMUN 57, 388–403.
[7] Duflou, J.R., Sutherland, J.W., Dornfeld, D., Herrmann, C., Jeswiet, J., Kara, S., Hauschild, M., Kellens, K., 2012. Towards energy and resource efficient manufacturing: A processes and systems approach. CIRP Annals-Manufacturing Technology 61, 587–609.
[8] Duque Ciceri, N., Gutowski, T., Garetti, M., 2010. A tool to estimate materials and manufacturing energy for a product, IEEE.
[9] Garretson, I.C., Eastwood, C.J., Eastwood, M.D., Haapala, K.R., 2014. A software tool for unit process-based sustainable manufacturing assessment of metal components and assemblies, in: ASME 2014 IDETC/CIE, ASME. pp. V004T06A047–V004T06A047.
[10] Heijungs, R., Henriksson, P.J., Guinée, J.B., 2017. Pre-calculated LCI systems with uncertainties cannot be used in comparative LCA. INT J LIFE CYCLE ASSESS 22, 461–461.
[11] Heilala, J., Vatanen, S., Tonteri, H., Montonen, J., Lind, S., Johansson, B., Stahre, J., 2008. Simulation-based sustainable manufacturing system design, in: Proceedings of the 40th Conference on Winter Simulation, Winter Simulation Conference. pp. 1922–1930.
[12] ISO/TS 14048:2002, 2002. Environmental management – Life cycle assessment – Data documentation format. ISO.
[13] Jiang, Z., Zhang, H., Sutherland, J.W., 2012. Development of an environmental performance assessment method for manufacturing process plans. INT J ADV MANUF TECH 58, 783–790.
[14] Kim, D.B., Shin, S.J., Shao, G., Brodsky, A., 2015. A decision-guidance framework for sustainability performance analysis of manufacturing processes. INT J ADV MANUF TECH 78, 1455–1471.
[15] Kuczenski, B., Marvuglia, A., Astudillo, M.F., Ingwersen, W.W., Satterfield, M.B., Evers, D.P., Koffler, C., Navarrete, T., Amor, B., Laurin, L., 2018. LCA capability roadmapproduct system model description and revision. INT J LIFE CYCLE ASSESS , 1–8.
[16] Kulkarni, A., Balasubramanian, D., Karsai, G., Narayanan, A., Denno, P., 2016. A domain specific language for model composition and verification of multidisciplinary models, in: Proceedings of the 2016 Annual Conference on Systems Engineering Researach, Huntsville, Alabama, USA.
[17] Mutel, C., 2017. Brightway: an open source framework for life cycle assessment. Journal of Open Source Software 12, 2.
[18] Overcash, M., Twomey, J., 2012. Unit process life cycle inventory (UPLCI)–a structured framework to complete product life cycle studies, in: Leveraging Technology for a Sustainable World. Springer, pp. 1–4.
[19] Rodríguez, M.T., Andrade, L.C., Bugallo, P.B., Long, J.C., 2011. Combining lct tools for the optimization of an industrial process: material and energy flow analysis and best available techniques. Journal of hazardous materials 192, 1705–1719.
[20] Rousseaux, P., Labouze, E., Suh, Y.J., Blanc, I., Gaveglia, V., Navarro, A., 2001. An overall assessment of life cycle inventory quality. INT J LIFE CYCLE ASSESS 6, 299.
[21] Steubing, B., Mutel, C., Suter, F., Hellweg, S., 2016. Streamlining scenario analysis and optimization of key choices in value chains using a modular LCA approach. INT J LIFE CYCLE ASSESS 21, 510–522.

Bernstein, William; Tamayo, Cesar; Lechevalier, David; Brundage, Michael.
"Incorporating unit manufacturing process models into life cycle assessment workflows."
Paper presented at 26th CIRP Conference on Life Cycle Engineering (LCE), West Lafayette, IN, United States. May 7, 2019 - May 9, 2019.

47th SME North American Manufacturing Research Conference, NAMRC 47, Pennsylvania, USA

Integrating A Dynamic Simulator and Advanced Process Control using the OPC-UA Standard

Hasan Latif[a,*], Guodong Shao[b], Binil Starly[a]

[a]*Department of Industrial & Systems Engineering, The NC State University, Raleigh, NC 27606, USA*
[b]*National Institute of Standards and Technology, Gaithersburg, MD 20899, USA*

Abstract

Insufficient interoperability has long been an issue on the factory floor, however, new technologies and standards are enabling production systems to become more agile and interoperable. A communication standard can, for example, make interoperation among different vendor-specific software and hardware tools in production systems easier and more reliable. In this paper, we share our research results and experience for the establishment of a connection between a dynamic simulator and an advanced process controller in a manufacturing system using OPC-UA. The OPC-UA communication protocol, which is middleware, acts as a common interface between these systems. We established the client and server for communication and defined an exchange data structure based on the OPC-UA standard for a control problem in a chemical process plant. The case study is a proof of concept of the OPC-UA standard implementation to support interoperability for different domains.

Keywords: OPC-UA; Simulation; Advanced Process Control; Integration; Automation.

1. Introduction

Manufacturing industries generate huge amounts of data, however, because it is difficult to effectively exchange data among the variety of manufacturing systems and applications on the factory floor, these data typically cannot be fully exploited. A McKinsey Global Institute report states that U.S. manufacturing can maintain competitiveness by applying an optimized, autonomous factory approach in a digitized and integrated value chain [14]. Achieving such digitization and integration within a manufacturing factory is, however, heavily dependent on plant floor data and communication techniques and common semantics. A free communication flow among different software and hardware systems, in turn, typically critically relies on the proper application of standards and proven methods [12]. Exploitation of advanced technologies through application of standards is one of the major challenges facing the manufacturing industry [4].
In recent years, there has emerged a significant shift in the interconnection of physical components on the manufacturing floor where transmitted information is used for control purposes [3]. Many quintessential requirements have been identified: ubiquitous connectivity, local intelligence, safety, self-organization, flexibility, massive data monitoring, and efficiency, to name a few [10] . The most common and crucial identified factor is efficient and reliable communication when dealing with heterogeneity and interoperability of various entities on a manufacturing floor. Advanced communication and information technologies can help achieve reliable, smooth, and robust integration between manufacturing levels through various physical media and protocols [21]. This paper reports a case study that integrates a simulator and controller via communication protocol: OPC UA. OPC UA is a sophisticated, scalable and flexible mechanism for establishing secure connections between clients and servers. This paper uses Tennessee Eastman problem to formulate the base problem. [5].
This section also provides a baseline for Digital Twin in terms of system integration, automation, and control. It establishes a standard-based communication between two different platforms on a manufacturing floor. The flexible and scalable communication approach can be used for similar manufacturing problems. The rest of the section is structured as follows: Section 2 discusses the related work, Section 3 provides the details

* Corresponding author. Tel.: +1-520-330-6058 ; fax: +1-919-515-5281.
E-mail address: hhlatif@ncsu.edu (Hasan Latif).

Latif, Hasan; Shao, Guodong; Starly, Binil.
"Integrating A Dynamic Simulator and Advanced Process Control using the OPC-UA Standard."
Paper presented at 47th SME North American Manufacturing Research Conference, NAMRC 47, Erie, PA, United States. June 10, 2019 - June 14, 2019.

of the case study implementation, Section 4 presents the findings, and Section 5 provides the conclusion and discussion.

2. Background

A common infrastructure model and communications framework can improve interoperability, enable more secure and efficient data transmission, and facilitate smart data usage. The research community has made significant efforts to introduce digital communications in control and field networks [7]. In this paper, we use the OPC-UA standard to enable communication and integration between the controller and the simulation of the Tennessee Eastman process. The reasons behind selecting OPC-UA over publish-subscribe technologies are multi-folds. It is platform independent, scalable, user friendly, and secure. OPC-UA is a completely new paradigm for systematic communication.

2.1. OPC-UA

Increasing demand for data exchange in a manufacturing plant requires better efficiency in communication networks [15]. As a result, newer advanced automation and control domains continue to emerge [1]. These new domains face a continuously increasing requirement of integration and interoperation. Therefore, standardized communication protocols are crucial for integrating manufacturing systems [11] [8]. OPC emerged as an automation standard primarily driven by automation vendors in process industry [19]. OPC defines a standard set of objects, interfaces, and methods to facilitate interoperability between control devices and systems. OPCs connectivity layer helps improve system interoperability.
In the 1990s, Microsoft introduced the Component Object Model (COM) and the Distributed COM (DCOM) interface standards. In 1995, Rockwell, Opto22, Intellution, and Fisher Rosemount developed a data-access standard based on COM and DCOM, and called OPC. Classical OPC include DA (Data Access), AE (Alarm & Events), HDA (Historical Data Access), and DX (Data Exchange). Each of these interfaces has a unique read and write command structure that impacts only one interface at the time. OPC-UA can be implemented on multiple platforms and no longer relies on COM/DCOM technologies [20].
The objective of the OPC-UA is to fulfill all the requirements for platform-independent system interfaces with versatile modeling capabilities that satisfy the needs of even complex systems. Independence of platform and scalability are necessary to facilitate the integration of OPC interfaces directly into a system that runs on various platforms. Access control and security are also crucial requirements because communication should be allowed through firewalls. The basic premise of OPC-UA is that the client can access small pieces of data without having to understand the entire complex model.
Therefore, it is widely accepted as an enabling technology for digital manufacturing [16]. So far, OPC-UA has been implemented by almost 20 different industrial sectors including tobacco, pharmaceuticals, and automation industries. These users have documented limitations including insufficient semantics, data models, dependence on COM/DCOM technology, inadequate security, and lack of implementation Application Programming Interfaces (API). This paper reports our effort to address these concerns by providing a case study and an implementation scenario.

2.2. Tennessee Eastman Problem

Tennessee Eastman (TE) is a well-known industry problem [5]. The original TE problem has a complex structure. It includes a three-unit operation: an exothermic, two-phase reactor; a flash separator; and a stripper. The TE problem contains 41 measured output variables and 12 manipulated variables. The TE problem has been solved with efficient algorithms using different modeling languages and tools. Downs & Vogel (1993) provided FORTRAN code of the model but did not publish the model equations. As a replacement, they provided a flow sheet, a steady-state material balance, and a qualitative description of the critical process characteristics. So, researchers who adopt the case need to make some assumptions to fulfill the missing information. In this case study, a simplified version of the TE problem has been adopted [18]. The simplified TE problem is in the steady state with a relatively modest structure. The details of the simplified TE problem will be discussed in the next Section.

3. Case Study

In this paper, we have adopted the simplified version of the TE process as the case study and performed the simulation and control modeling. We model the controller and the TE problem simulation using two different applications between which information must constantly flow. OPC-UA acts as middleware between the applications. The scenario is pragmatic and can be reused for other similar real-world cases. This case study serves as a prime example of demonstrating how OPC-UA is implemented for data communication within a plant.
Figure 1 illustrates the overview of information flow of the simplified TE process. First, we have derived an optimization problem from the simplified TE case. Then, we identify the optimal parameters for controllers. These optimal parameters, as control set points, are then sent to the process simulator via OPC-UA. Simulator sends the feedback to the controller at regular intervals. With this feedback, controller adjusts the new optimal parameters and send them back to process simulator. Therefore, a continuous and effective communication takes place at regular interval.

3.1. Simplified Tennessee Eastman Problem

As shown in Figure 2, the simplified TE problem includes a combined reactor and separator vessel. The model has two input flows (Feed 1 and Feed 2) and two output flows (Feed 3 and Feed 4). Feed 1 admits gas compounds A and C into the reactor while pure A is used to control the ratio between A and

Latif, Hasan; Shao, Guodong; Starly, Binil.
"Integrating A Dynamic Simulator and Advanced Process Control using the OPC-UA Standard."
Paper presented at 47th SME North American Manufacturing Research Conference, NAMRC 47, Erie, PA, United States. June 10, 2019 - June 14, 2019.

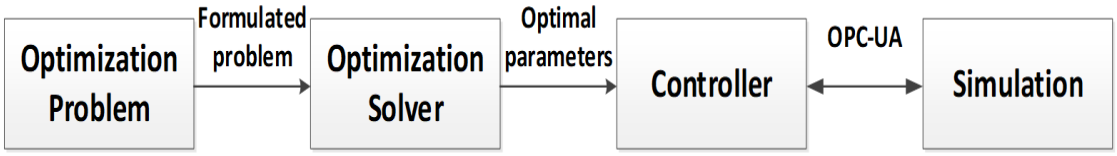


Fig. 1. Information Flow between modules of the case study

C through Feed 2. Product D, a liquid, exits through Feed 4, while the purge vapor flows out through Feed 3. In summary, the inputs of the system are A and C, and the outputs are D and the vapor purge as seen in Equation 1.

$$A + C = D + Purge \tag{1}$$

3.2. Optimization Problem Formulation

The optimization problem is used to derive the optimal parameter values to serve as control set points. The optimization problem has been formulated based on the model developed by Ricker (1993), which assumes that the plant is at the steady state. The optimization objective of the problem is to minimize the instantaneous cost of producing a given amount of product D per hour, which depends on three user-provided input parameters: the product flow rate in kmol per hour, the cost per kmol of A, and the cost per kmol of C. The optimization result includes optimal values for six parameters that allow users to enact the most cost-effective setup. The parameters manipulated to achieve minimum cost are the valve positions (as a percentage open) of Feeds 1 to 3 as well as the total pressure of the system. From these values, the valve position of Feed 4 can also be calculated. These five variables, as well as the instantaneous cost, are returned after the optimization execution. The mathematical model is described below:

$$\begin{aligned} Minimize\ C = & 1/F_4 * [C_A(y_{A1}\chi_1 F_{1max} + \chi_2 F_{2max} - F_4) \\ & + C_C(y_{C1}\chi_1 F_{1max} - F_4)] \end{aligned} \tag{2}$$

such that

$$\begin{aligned} & k_0\left(\frac{P}{\chi_3 C_{v3}\sqrt{P-100}}\right)^{1.6} * (y_{A1}\chi_1 F_{1max} + \chi_2 F_{2max} - F_4)^{1.2} \\ & * (y_{C1}\chi_1 F_{1max} - F_4)^{0.4} - F_4 \ \leq \ 0 \\ & \text{and} \\ & y_{C1}\chi_1 F_{1max} \ \geq \ 0.8(y_{A1}\chi_1 F_{1max} + \chi_2 F_{2max}) \end{aligned} \tag{3}$$

where,
χ_1 = Feed 1 valve position (%, expressed as decimal)
χ_2 = Feed 2 valve position (%, expressed as decimal)
χ_3 = Purge valve position (%, expressed as decimal)
$\chi_4 = \frac{P}{\chi_3 C_{v3}\sqrt{P-100}}$
= Product valve position (%, expressed as decimal)
P = Total pressure of system (kPa)
F_4 = Product flow (kmol/h)
C_A = Cost of A ($/kmol)
C_C = Cost of C ($/kmol)
y_{A1}= Concentration of A in Feed 1 (%, expressed as decimal)
y_{C1} = Concentration of C in Feed 1 (%, expressed as decimal)
F_{1max} = Maximum flowrate of Feed 1 (kmol/h)
F_{2max} = Maximum flowrate of Feed 2 (kmol/h)
k_0= Constant value associated with reaction

Equation 2 represents the relationship between the reaction rate of the system and the product flow rate based on the time-based equations from the model. Since the problem was assumed to be in steady state, equation 2 was derived by setting Rickers state equations (1) through (4) equal to zero [18]. The cost equation naturally favors A, so equation 3 ensures that an ideal ratio between A and C is maintained. Table 1 lists the variables and their descriptions. The variables are sorted into three categories: output variables, input parameters, and nominal values. An optimal value is assigned the output variables by the optimization solver, input parameters, as constants, are provided by the user, nominal values are taken from Table 1 [18].

After executing the optimization, the optimal and target values (feed valves 1, 2, 3, and 4) are derived. These target values are used by the controller as set points values.

3.3. OPC-UA Client Development: Advanced Process Control

Process control plays a vital role ensuring conformity to process rules and protecting the process environment. Real-time optimization (RTO) can be deployed in a controller to determine the optimum control set points for the current operating conditions and constraints. The operating constraints for a plant are identified as part of the process design. During plant operations, the optimum operating conditions can change regularly owing to product throughput, process disturbances, by-product as wastes, and economic evaluations. Therefore, it is profitable to recalculate the optimum operating conditions on a regularly.

Latif, Hasan; Shao, Guodong; Starly, Binil.
"Integrating A Dynamic Simulator and Advanced Process Control using the OPC-UA Standard."
Paper presented at 47th SME North American Manufacturing Research Conference, NAMRC 47, Erie, PA, United States. June 10, 2019 - June 14, 2019.

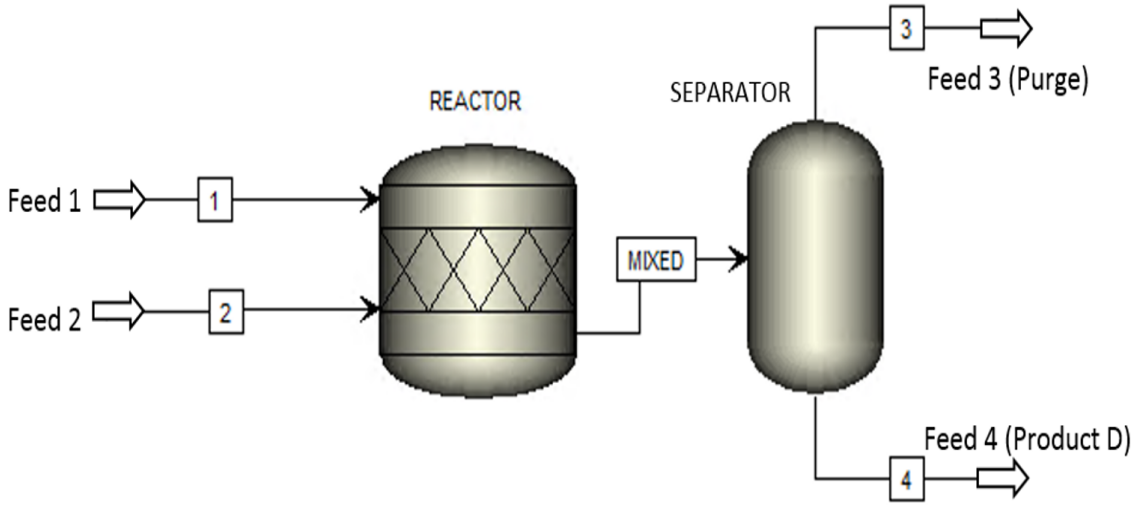


Fig. 2. Process Schematic of the Simplified TE Problem

Table 1. Summary of variables and nominal operating conditions

Output Variable	Set Value	Description	Units
χ_1	0.609533	Feed 1 valve position	(%)
χ_2	0.250223	Feed 2 valve position	(%)
χ_3	0.392578	Feed 3 valve position	(%)
χ_4	0.470302	Feed 4 valve position	(%)
P	2700	Total system pressure	kPa
C	0.2415	Instantaneous cost	$/kmol
Manipulated Variable	**Set Value**	**Description**	**Units**
F_4	100	Product flowrate	Kmol/hour
C_A	2.206	Cost of A	$/kmol
C_C	6.177	Cost of C	$/kmol
Constants	**Set Value**	**Description**	**Units**
y_{A1}	0.485	Concentration of A in Feed 1	(%)
y_{C1}	0.510	Concentration of C in Feed 1	(%)
F_{1max}	330.46	Max flow rate of Feed 1	Kmol/hour
F_{2max}	22.46	Max flow rate of Feed 2	Kmol/hour
k_0	0.00117	Constant for assumed isothermic reaction	–

In this paper, a model predictive control (MPC) is designed to control the TE process simulation. A predictive model controller is part of a multi-level control hierarchy in modern processing plants [21]. We use Aspen DMC3 to develop the MPC controller [9].

Three different types of variables are used: manipulated (MV), controlled (CV), and disturbance variables (DV). The three manipulated variables are three valve positions: U_1, U_2, and U_3 respectively. The three controlled variables are product flowrate F_4, pressure (P), and product A in the by-product Y_{A3}.

The relationship between controlled variables and manipulated variables are adapted from Ricker (1993) [18]. The connections are derived in transfer function format from a state space model of the TE problem.

$$y = \begin{bmatrix} F_4 \\ P \\ y_{A3} \end{bmatrix} = Gu = \begin{bmatrix} g_{11} & 0 & 0 \\ g_{21} & 0 & g_{23} \\ 0 & g_{32} & 0 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} \tag{4}$$

$$g_{11} = \frac{1.7}{0.75s + 1} \tag{5}$$

$$g_{21} = \frac{45(5.67s + 1)}{2.5s^2 + 10.25s + 1} \tag{6}$$

$$g_{23} = \frac{-23.81s - 2.086}{s^2 + 7.874s + 0.1915} \tag{7}$$

$$g_{32} = \frac{1.5}{10s + 1} e^{-0.1s} \tag{8}$$

Latif, Hasan; Shao, Guodong; Starly, Binil.
"Integrating A Dynamic Simulator and Advanced Process Control using the OPC-UA Standard."
Paper presented at 47th SME North American Manufacturing Research Conference, NAMRC 47, Erie, PA, United States. June 10, 2019 - June 14, 2019.

The constraints of the model are given as below.

- Pressure (P) has an upper bound (3000 Kpa).
- A in the purge $yA3$ has a range (0.429 ¡ $yA3$ ¡ 0.886).
- Product flow F_4 has a set point (100 Kmol/hr).
- All threee manipulated variables are unconstrained.

Using the transfer functions, library models are created in Aspen DMC3. Different types of state space models can be stored in a library. These library models can be reused to establish the relation between manipulated and controlled variables. For instance, the first order transfer functions library model formula is $\frac{K}{T*s+1}e^{-D*s}$

where T = Time Constant, D = Delay, K = Gain.
In this problem, transferring g32 to model library provides T = 10 mins; K = 1.5; D = 6 sec.
Aspen DMC3 provides a visual representation of the library model as well. The graph of transfer function g_{32} is given in Figure 3.

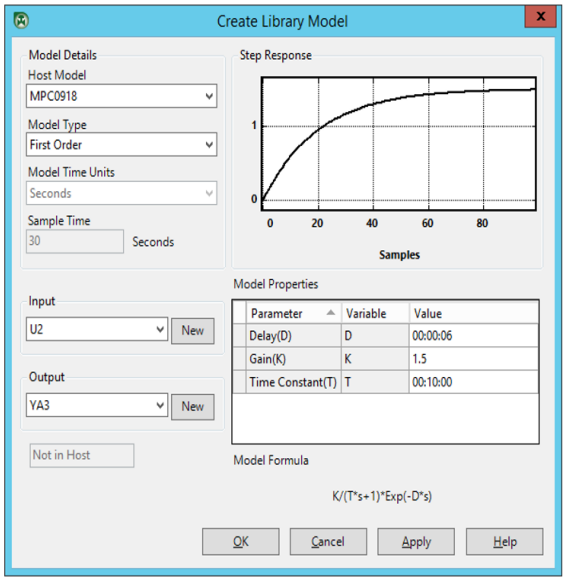


Fig. 3. Transfer Function Graph of a Model Predictive Controller (MPC)

After storing all the transfer functions in the library model, a master model is prepared. After being simulated offline, the controller is ready to deploy.

3.4. OPC-UA Server Development: Simulation

A Modelica model has been developed [13] to simulate the dynamic behavior of the simplified TE process, and it is based on the mathematical description provided by Ricker (1993). This Modelica model library includes the model of the open loop plant. It consists of a model named Reactor, a connector designated pCon, models for setting the boundary conditions, and two models describing the input and output source valves. The Reactor model represents the processing unit that combines the behavior of the reactor and the separator. These models have been used to compose the ReactorOpenLoop model, a model describing the behavior of the open loop plant. The model library is written in Modelica 3.3 and has been tested using Dymola 2018 and OpenModelica 1.11.0 64 bits under Windows 2010. The model has been used for the ISO 15746 standard implementation.

3.5. Communication Protocol: OPC-UA

The Modelica simulation, discussed in previous subsection, is acting as an OPC server. The controller, designed in Aspen DMC3, is acting as an OPC client.
To make a successful OPC-UA connection, OPC client and server need to communicate via nodes. The OPC server needs to identify the nodes and read data successfully. To setup an OPC client using Aspen DMC3, the Cim-IO interface manager first needs to be started. The CIM-IO interface is a communication interface that provides a communication standard for interfacing with various AspenTech products like InfoPlus.21 and third-party software such as Modbus, OPC servers. Through DMC3s CIM-IO interface manager, the OPC-UA interface gets active and ready to communicate with the server.
Next, the OPC-UA client requires connecting to the OPC-UA server via nodes. The nodes addresses are provided in the modeling. Then the OPC-UA connection via nodes is tested and deployed. Figure 4 is a screen capture that shows the variable names and types, as well as the node address assignment.
After a successful OPC-UA server/client connection, the controller MPC1 is deployed and starts running. In the Aspen Web Interface module, the feedback from controller and simulation is observable. The history, data exchange information, and controller application can be seen and changed from this module according to the user need.

The optimization execution result provides the controller with target set values for feed valve positions. The controller uses the constraints to acquire real-time feed valve positions and communicates with the simulation through the OPC-UA messaging protocol. The simulation also provides feedback to the controller via OPC-UA and the controller acts as a check and balance element in the simulation by providing the next set of real-time feed valve positions. Overall, this case study is an implementation of a standard-based communication protocol in the manufacturing domain. The approach can be applied to similar problems in the plant to enable real-time communication between different enterprise levels. Automotive, medical device, consumer electronics, aerospace & defense industry can adapt the technology and march towards digital manufacturing.

4. Lessons Learned

Even though in this case study, the OPC-UA has been successfully implemented between a controller and a simulator of

Latif, Hasan; Shao, Guodong; Starly, Binil.
"Integrating A Dynamic Simulator and Advanced Process Control using the OPC-UA Standard."
Paper presented at 47th SME North American Manufacturing Research Conference, NAMRC 47, Erie, PA, United States. June 10, 2019 - June 14, 2019.

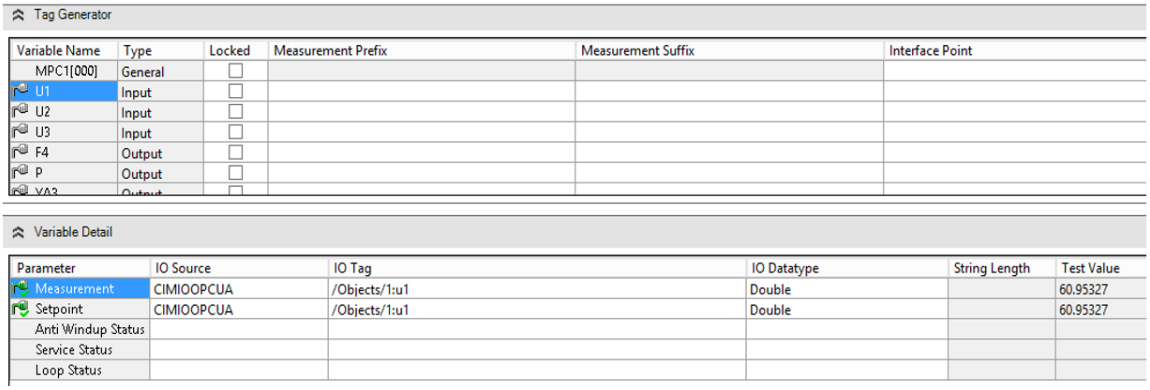


Fig. 4. Test and Deployment of Node Address

the simplified TE problem, OPC-UA implementation requires some complicated procedures. Multiple challenges need to be addressed in a proficient manner to have a smooth OPC-UA implementation. We have identified the following such major challenges based on our experience.

- The challenge of selecting OPC-UA server/client enabled applications and specification of a well-defined architectures: not all the control and simulation applications are OPC-UA enabled, so effort is needed to select an application that is not only capable of modeling the problem (e.g., control) but also establishing a server or a client. Also, OPC-UA has a large set of specifications. It is difficult to assess and estimate the project effort and development time in the beginning of the project. The existing physical system has a limited capacity that can sometimes hinder additional functionality. For instance, the existing system has only a limited amount of RAM space or processor clockwork speed available for additional OPC-UA accommodation. As OPC-UA memory utilization increases, it poses a threat to the existing infrastructure to crash.
- Because OPC-UA connects a multitude of applications across firewalls and networks, server security becomes a concern. Like many other message protocol systems, OPC-UA uses authentication, authorization, and encryption via an address space concept. OPC-UA address space provides a standard way for servers to represent objects to clients. It defines objects in terms of variables and methods. The elements of a model are represented in the address space as nodes that are assigned to a node class, e.g., objects, variables, and methods. On the other hand, the software components have different levels of maturity for creating the address space model. Some of these software components provide a graphical user interface (GUI) to model address space and to add nodes and references, while others do not. The GUI generates code to establish OPC-UA connection between servers and clients. With a GUI, therefore, it will be easier for building up the server/client connection. Although the node address can be identified from literature review, experiments, or manuals, none of these methods is very intuitive. For instance, in this case study, we have identified the OPC-UA server node address via a reference and the OPC-UA client node address by setting up an additional test server. With the help of the test server, the connection is established with the OPC-UA client and the node address is captured through the connection details.
- An OPC-UA server contains sets of services that are used by the clients. All the OPC-UA functionalities exist in these services, which have a request and response message. Services are defined in the OPC-UA standard and the user cannot change them. For instance, while implementing the OPC-UA communication between server and client, it is very important to ensure that each follows the same (32 bit/64bit) communication bits. Users cannot modify any of the system-defined architecture.
- Security protocol is pre-defined in OPC-UA. OPC-UA claims all the required security features are built in to minimize the efforts from the developers [2]. In this case study, certificate authority is used as a security measure to ensure data protection. A Certificate Authority creates and verifies certificates. It also adds a digital signature to the Certificate to confirm user identity. OPC UA applies different security tiers. It is found that security tiers largely depends on the application platform. User has little to none control over choosing the security tier once the application platform is chosen.
- There is no way to measure the performance of the OPC-UA connection currently. At any point of OPC-UA establishment, it is hard to understand the level of reliability and quality of connection. Moreover, OPC-UA cannot determine data quality which impacts the performance as well. Should it become possible to quantify or approximate the performance of the OPC-UA connection, different methods could apply, e.g., memory rearrangement, structural reallocation, existing firmware upgrades.

Latif, Hasan; Shao, Guodong; Starly, Binil.
"Integrating A Dynamic Simulator and Advanced Process Control using the OPC-UA Standard."
Paper presented at 47th SME North American Manufacturing Research Conference, NAMRC 47, Erie, PA, United States. June 10, 2019 - June 14, 2019.

- Finally, to accomplish a complete semantics interoperability, OPC-UA alone is not sufficient because it only enables syntactic interoperability between clients and servers. There is a great need for semantics to support analytics and scalability across various application from different vendors.

Overall, these key challenges contribute to the cost of OPC-UA implementation and create additional uncertainties.

5. Discussion and Conclusion

Interoperability is a very critical issue that manfuacturers have to deal with. Communication standards such as OPC-UA make it possible to have efficient and reliable information exchange between enterprise levels. By integrating with other manufacturing interoperabilty standards with semantics such as MTConnect e.g., MTConnect-OPC-UA cmpanion specification, OPC-UA could become an important piece of in semantic interoperability for industrial applications. Even though OPC UA provides the largest eco-system for industrial operability, OPC foundation has unveiled a new version of OPC-UA called OPC-UA PubSub [6][17]. PubSub enables the use of OPC UA directly over the Internet by utilizing popular data transports like MQTT and AMQP. At the same time, it retains key OPC UA end-to-end security and standardized data modeling advantages.

This paper presents an approach for implementing OPC-UA as middleware between different manufacturing systems. The integration of process simulation and advanced process control from two different application environments has not been done before. The case study identified the implementation requirements for the applied problem, standards, and technologies. The feasibility of scenario is also verified in the case study. Valuable lessons learned have been discussed.

However, in this work, the OPC-UA technology was tested with only a small stream of data in a laboratory environment. In a real-work application, enormous amounts of data have to be transferred and exchanged, which may complicate the implementation. The performance of the OPC-UA data exchange needs to be studied more closely. There are several ways that this work could be continued in the future. A manufacturing case study in which more vendors products are involved for integration; a real process that generates more complex data and more realistic amounts of data could be used to replace the process simulator. In addition, the methodology and scenario of the implementation could be enhanced. More functionalities could be implemented within the system.

Disclaimer

No approval or endorsement of any commercial product by NIST is intended or implied. Certain commercial software systems are identified in this paper to facilitate understanding. Such identification does not imply that these software systems are necessarily the best available for the purpose.

References

[1] Alcaraz, C., Roman, R., Najera, P., & Lopez, J. (2013). Security of industrial sensor network-based remote substations in the context of the internet of things. Ad Hoc Networks, 11(3), 1091-1104.

[2] Armstrong, R. & Hunkar, P. (2010). The OPC UA security model for administrators. White paper V1.00, OPC Foundation. Link: https://opcfoundation.org/wp-content/uploads/2014/05/OPC-UA_Security_Model_for_Administrators_V1.00.pdf

[3] Bradley, D., Russell, D., Ferguson, I., Isaacs, J., MacLeod, A., & White, R. (2015). The Internet of ThingsThe future or the end of mechatronics. Mechatronics, 27, 57-74.

[4] Cwikla, G., & Foit, K. (2017). Problems of integration of a manufacturing system with the business area of a company on the example of the Integrated Manufacturing Systems Laboratory. In MATEC Web of Conferences (Vol. 94, p. 06004). EDP Sciences.

[5] Downs, J. J., & Vogel, E. F. (1993). A plant-wide industrial process control problem. Computers & chemical engineering, 17(3), 245-255.

[6] Drahos, P., Kucera, E., Haffner, O., & Klimo, I. (2018, January). Trends in industrial communication and OPC UA. In 2018 Cybernetics & Informatics (K&I) (pp. 1-5). IEEE.

[7] Espi-Beltran, J. V., Gilart-Iglesias, V., & Ruiz-Fernandez, D. (2017). Enabling distributed manufacturing resources through SOA: The REST approach. Robotics and Computer-Integrated Manufacturing, 46, 156-165.

[8] Fruhwirth, T., Pauker, F., Fernbach, A., Ayatollah, I., Kastner, W., & Kittl, B. (2015, November). Guarded state machines in OPC UA. In Industrial Electronics Society, IECON 2015-41st Annual Conference of the IEEE (pp. 004187-004192). IEEE.

[9] Golightly, R. 2018. The Aspen DMC3 Difference. White paper.

[10] Gonzalez, I., Calderon, A. J., Barragan, A. J., & Andujar, J. M. (2017). Integration of Sensors, Controllers and Instruments Using a Novel OPC Architecture. Sensors, 17(7), 1512.

[11] Hernandez, L., Baladron, C., Aguiar, J. M., Calavia, L., Carro, B., Sanchez-Esguevillas, A., & Gomez, J. (2012). A study of the relationship between weather variables and electric power demand inside a smart grid/smart world framework. Sensors, 12(9), 11571-11591.

[12] Lu, Y., Morris, K. C., Frechette, S. (2016). Current Standards Landscape for Smart Manufacturing Systems. NISTIR 8107.

[13] Martin-Villalba, C., Urquia, A., & Shao, G. (2018). Implementations of the Tennessee Eastman Process in Modelica. In Proceeding of the 9th Vienna International Conference on Mathematical Modeling. Elsevier, IFAC-PapersOnLine. Volume 51. Issue 2. Pages 619-624.

[14] Ramaswamy, Sree, J. Manyinka, G. Pinkus, K. George, J. Law, T. Gambell, and A. Serafino. "Making it in America: Revitalizing US manufacturing." New York, NY: McKinsey Global Institute. Google Scholar (2017).

[15] Mejias, A., Herrera, R. S., Marquez, M. A., Calderon, A. J., Gonzalez, I., & Andujar, J. M. (2017). Easy Handling of Sensors and Actuators over TCP/IP Networks by Open Source Hardware/Software. Sensors, 17(1), 94.

[16] Pauker, F., Frhwirth, T., Kittl, B., & Kastner, W. (2016). A systematic approach to OPC UA information model design. Procedia CIRP, 57, 321-326.

[17] OPC Foundation Press Release, March, 2018. Link: https://opcfoundation.org/news/press-releases/opc-foundation-announces-opc-ua-pubsub-release-important-extension-opc-ua-communication-platform/

[18] Ricker, N. L. (1993). Model predictive control of a continuous, nonlinear, two-phase reactor. Journal of Process Control, 3(2), 109-123.

[19] Rohjans, S., Piech, K., & Lehnhoff, S. (2013, November). UML-based modeling of OPC UA address spaces for power systems. In Intelligent Energy Systems (IWIES), 2013 IEEE International Workshop on (pp. 209-214). IEEE.

[20] Toivanen, S. (2016). Implementation of performance assessment tool for multivariable, model-predictive controller. MS Thesis, Tampere University of Technology.

[21] Wagner, T. (2003, September). Applying agents for engineering of industrial automation systems. In German Conference on Multiagent System Technologies (pp. 62-73). Springer, Berlin, Heidelberg.

Latif, Hasan; Shao, Guodong; Starly, Binil.
"Integrating A Dynamic Simulator and Advanced Process Control using the OPC-UA Standard."
Paper presented at 47th SME North American Manufacturing Research Conference, NAMRC 47, Erie, PA, United States. June 10, 2019 - June 14, 2019.

KEY IMPLEMENTATION CHALLENGES AND CROSSCUTTING RESEARCH THEMES FOR DEVELOPING IMMEDIATE OCCUPANCY PERFORMANCE OBJECTIVES

Siamak Sattar, Christopher L. Segura Jr., Katherine J. Johnson, Therese P. McAllister,
and Steven L. McCabe
National Institute of Standards and Technology (NIST)
Gaithersburg, MD USA

Abstract

New and existing buildings and supporting infrastructure may sustain extensive damage during natural hazard events such that building functions are degraded or lost. Widespread building damage across a community can have severe social and economic impacts. The U.S. Senate tasked NIST with identifying research needs and implementation activities to develop multi-hazard immediate occupancy (IO) performance objectives for commercial and residential buildings. With input from subject matter experts and stakeholders participating in a national workshop, NIST developed a report that describes research areas and implementation activities to fulfill the Congressional mandate. The content of the report is organized around four topic areas: enhancing building design, addressing community considerations, ascertaining social and economic issues, and identifying acceptance and adoption considerations that require further reflection in the process of developing the new IO performance objective. This paper discusses crosscutting themes that apply to all four topic areas and will need to be addressed to advance research and implementation activities for IO performance. These crosscutting themes define activities that are vital to the development of research tools, design standards, and educational tools needed to study the impacts of, and design for, IO performance. The paper also highlights key challenges in adoption and implementation of IO performance objectives; these challenges focus mainly on social, economic, and policy related issues that can support the successful adoption of IO objectives by the public.

Introduction

The U.S. Senate requested that the National Institute of Standards and Technology (NIST) create a report to plan for improvements to "the resiliency of buildings, homes, and infrastructure" for the American public (U.S. Senate, 2016). The congressional mandate was motivated by the reality that "current building codes often do not provide the necessary protection against natural hazards, particularly with regard to enabling immediate occupancy after a significant earthquake, hurricane, tornado, flood, or other natural disaster" (U.S. Senate, 2016). Communities, owners, and residents should benefit from buildings that are more resilient to natural hazard events to avoid lengthy and costly repairs or rebuilding, as well as minimizing the need for long-term evacuation of building occupants. Thus, the Senate directed NIST to identify engineering principles, research, and implementation activities needed for a new "safety building performance objective for commercial and residential properties" (U.S. Senate, 2016). In response to the congressional mandate, NIST developed a report identifying the research needs and implementation activities required to develop IO performance objectives (Sattar et al., 2018). This report was developed through a collaborative process with a steering committee of subject matter experts and a national expert stakeholder workshop hosted by NIST. In the NIST report, immediate occupancy (IO) performance is considered as the building's condition after a hazard event where damage to the building's structural system is controlled, limited, and repairable while the building remains safe to occupy. The building's ability to function at full or minimally reduced capacity is also affected by the functionality of the non-structural systems of the building (e.g., building envelope, equipment, interior utilities), as well as the infrastructure that connects the building to its surrounding community. The term IO is used for general reference to a

Sattar, Siamak; Segura, Christopher; Johnson, Katherine; McAllister, Therese; McCabe, Steven.
"KEY IMPLEMENTATION CHALLENGES AND CROSSCUTTING RESEARCH THEMES FOR DEVELOPING IMMEDIATE OCCUPANCY PERFORMANCE OBJECTIVES."
Paper presented at 17th U.S.-Japan-New Zealand Workshop on the Improvement of Structural Engineering and Resilience, Queenstown, New Zealand. November 12, 2018 - November 14, 2018.

potential range of functional levels for consistency with the congressional language. The role of lifelines in supporting the operation of functional buildings is acknowledged, but not addressed in detail in the NIST report. The NIST report covers improvements to building design, as well as community, economic and social, and adoption and acceptance considerations. This paper highlights crosscutting research needs and key implementation challenges to IO performance objective development identified in the NIST report.

Motivation

By ensuring continuing access to housing and resumption of local businesses following a hazard event, communities can use IO buildings to mitigate and recover from natural hazards and to reduce vulnerability and long-term negative consequences. Geographic regions in the United States face a unique combination of natural hazards. As shown in Fig. 1, significant weather and climate disasters in the continental U.S. in 2017 were widespread. These weather events cause extensive damage and disruption to buildings, loss of life, injury, property damage, displacement of residents and businesses, and have long-lasting economic and social effects that impact local communities and the spirit of the nation (NOAA, 2018). It is reported that 2017 was the costliest year for weather and climate events in the United States, with the U.S. incurring $306B in natural hazard damages (Mooney and Dennis, 2018). It is important to note that earthquakes are not included in NOAA's reports of weather and climate disasters (Fig. 1), and that the U.S. has not experienced a major damaging earthquake since 1994. However, the Federal Emergency Management Agency (FEMA) estimates the annualized cost of damage to U.S. building stock from earthquakes to be $6.1B per year (FEMA, 2017), and this should be included to more accurately assess risk.

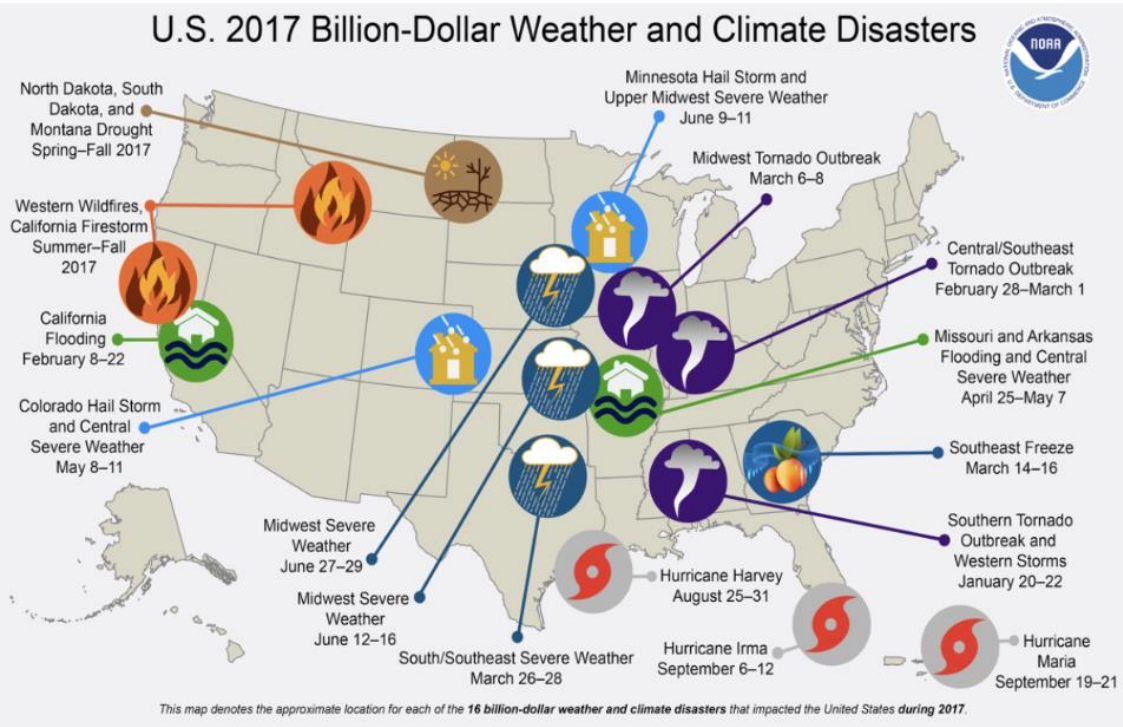


Figure 1. Regional variation of significant 2017 weather and climate disasters (source: NOAA, 2018)

The economic costs borne by individuals, governments, and insurance companies, as a result of natural hazard events, are substantial and provide an important measure of disaster impacts. Similar to reports for the U.S., 2017 was the costliest year for weather and climate disasters globally, with 710 recorded natural disasters (Munich Re, 2018). While 2017 may seem anomalous, it is representative of an increasing trend in occurrence of natural hazard events, with five of the past six years breaching the 600-event mark globally (Munich Re, 2018).

Development of new IO performance objectives to improve the resilience of buildings can help reduce damage and losses across all types of natural hazards, whether geologic or climatic and frequent or infrequent. In modern buildings, loss of life and structural collapse from natural hazard events are infrequent. The goal of building codes is to protect lives by reducing the likelihood of structural collapse for a design-level event (as defined in the codes), and to provide some level of property protection. However, societal needs are quickly outpacing this performance goal. A new performance objective in building codes would improve the performance of buildings and infrastructure, so that they are less likely to be negatively impacted and more likely to maintain functionality or regain it quickly. IO performance objectives could serve to reduce short- and long-term population displacement, adverse health effects, and disruption to communities caused by impairments to government, schools, and businesses.

Areas for Research and Action

The research and implementation activities required to develop IO performance objectives are organized around four main topics:

1) *Building design*: includes advances related to designing or retrofitting an individual building to meet IO performance objectives and changes to building code provisions;

2) *Community considerations*: discusses the resilience context for the role of buildings in community physical, social, and economic systems before and after hazard events;

3) *Economic and social considerations*: addresses feasibility of implementing IO performance and the potential impacts that improved building performance may have on social and economic systems;

4) *Acceptance and adoption considerations*: addresses activities required for effective implementation of IO performance by stakeholder communities, including state and local government officials, engineers, architects, urban planners, developers, building owners, and building occupants.

These topics were developed through an iterative process that included a literature review and subject-matter expert input from the steering committee members and workshop participants. The research and implementation needs associated with each of these topics are discussed in Sattar et al. (2018).

Crosscutting Research Needs to Develop Immediate Occupancy

Several crosscutting issues were identified that address research needs and implementation activities pertinent to all four of the main topics. These issues include needs to develop research tools for studying the impacts of IO performance objectives, as well as to develop new guidelines, standards, and educational tools to support the implementation of IO. Crosscutting issues are organized according to the following six categories:

- *Data* – Datasets on building performance and community impacts to support the development of research and implementation tools for IO
- *Relationships and Dependencies* – Characterization of relationships that describe the factors that influence building functionality and the interaction of a building or building cluster with the surrounding community
- *Predictive Analytical Models* – Science-based models to study the impacts of IO performance objectives across multiple spatial and temporal scales
- *Metrics and Tools* – Mechanisms to evaluate the anticipated performance of buildings and to assess community functions in relation to IO performance desires
- *IO Guidance Documents and Design Standards* – Criteria used by architects, engineers, and community decision-makers to assess IO performance
- *Education, Outreach, and Training* - IO-specific competency and qualification programs

Data. To assist with the development of analytical tools, decision support tools, and stakeholder communication tools, broad datasets reflecting the performance of building systems, social systems, and economic systems in the pre-hazard, post-hazard, and recovery time periods are needed. Existing field reconnaissance data, laboratory data, and results of analytical studies should be consolidated so that they are readily available to researchers and to enable an assessment of data collection methods. Existing reconnaissance data and laboratory data should be evaluated against analytical modeling needs to identify shortcomings in data to support the development and validation of analytical models.

Standardized reconnaissance data collection protocols need to be established to ensure datasets are comprehensive and capture critical information needed to calibrate and validate analytical models. Data collection methods should emphasize the need for consistency in terms of the types, quantity, and timing of data collected. These standardized methods should also emphasize the need to collect data for buildings of various ages and construction types with a range of damage levels. Furthermore, prioritization should be placed on continued, periodic evaluation of the recovery of functionality for damaged buildings and the recovery process for impacted communities in the post-hazard time period. New data collection technologies are needed for monitoring and assessing the performance of IO buildings and to develop models expressing the relationship between building damage levels and functionality levels.

Relationships and Dependencies. A common theme across the four topic areas is the need to develop models that communicate the relationships and dependencies between functional levels, damage and recovery levels, and the effects thereof on populations, social and economic systems, and communities. In the context of an individual building, it is important to better understand how the functional level of a building is impacted by aging and periodic natural hazard events during the building's lifecycle. These relationships should express functional levels with respect to the entire building's structural and nonstructural systems, as well as infrastructure services. Such relationships should describe the short-term impacts of interruption to community infrastructure services (e.g., water and power) and reduced levels of building functionality when temporary backup services (e.g., generators and water tanks) are used. The impact of maintenance, repair, and retrofit technologies should also be considered.

On a community level, the relationships and dependencies between community infrastructure services, individual buildings, and building clusters need to be better understood. The relationships should incorporate redundancies within community systems to provide a broad description of the effects of damage on the functionality and recovery of the physical, social, and economic systems of the community.

Predictive Analytical Models. Analytical models are needed to better understand interactions between the complex systems of a community and to study the direct and indirect effects of interventions, including the introduction of IO buildings into a community. The models should enable the prediction of performance on multiple spatial scales ranging from the individual systems of a building, to clusters of buildings that support particular community functions (e.g., healthcare, education, business, or governance), to the community scale involving all building clusters within the community. The models should also address multiple temporal scales ranging from several days to multiple decades for the pre-hazard, post-hazard, and recovery time periods. The analytical models should incorporate the relationships, described above, that address the impacts of damage, maintenance, repair, and retrofit strategies on functional levels. These tools will allow for the assessment of the integrated performance of a community's physical, social, and economic systems, including the dependencies among these systems.

Metrics and Tools. Performance metrics and analysis tools are needed to evaluate the anticipated performance of buildings designed for IO objectives. Performance metrics should describe the desired goals for building damage and functionality, community recovery, and social and economic well-being. Analysis tools should provide the means to analytically evaluate the ability of a building's systems to meet desired IO performance objectives and to assess existing community functions in relation to IO building performance desires.

IO Guidance Documents and Design Standards. Guidance documents and design standards are needed to support the implementation of IO performance objectives. These tools will articulate the technical evaluation criteria for a building and its systems relative to desired levels of building functionality and design level hazards. The design tools should contain guidance on identifying buildings for IO objectives, by considering their role in the community and their impact on social and economic systems.

Education, Outreach, and Training. Recruiting and maintaining a workforce knowledgeable about IO performance objectives and implementation methods will be crucial to ensure a common understanding across professions. For the engineering and architectural fields, designing buildings to IO performance objectives would be a notable shift from current practice. It may require an IO-specific set of competencies, and a licensure or accreditation program for designers, contractors, and code officials. This could have widespread implications on undergraduate and graduate curricula and future workforce recruitment and retention. In addition to the technical workforce necessary to design and construct IO buildings, code officials will need to be trained to enforce the design standards and ensure buildings are constructed to code and can meet IO performance objectives.

Building owners and community leaders need education on hazard risks, costs and benefits, and best practices. Additionally, opportunities for diverse sets of stakeholders to interact and communicate in a group setting, such as community workshops, should also be explored. Examples of these stakeholders include financial institutions, insurance companies, foundations, federal and state governments, business, utilities, commercial building owners, and homeowners.

Key Implementation Challenges to Develop Immediate Occupancy Performance Objectives

Research and implementation needs for developing IO building performance objectives is more than a technical problem of how to design and construct buildings that are more resilient to natural hazards. In addition to the four technical research topics discussed earlier, there are multiple complex social, economic, and policy challenges that should also be addressed to ensure successful adoption of IO performance objectives. The challenge of achieving IO performance is just as much a social and economic matter as it is a technical one. The key implementation challenges described below are outside the scope of the

crosscutting research topics or implementation activities discussed earlier and would require coordinated and cooperative work over time and across sectors.

Motivating Action. While communities often reflect on desired building performance in the wake of a natural hazard event, a key barrier to adoption and acceptance of an IO performance objectives is motivating the community to invest in improved building performance in advance of hazard events. There is a general expectation that current building codes and regulations protect against damage or loss of functionality from hazard events. In reality, building codes are primarily designed to safeguard lives and only provide some degree of property protection. Shifting public expectations to IO performance and functionality, and ensuring those objectives are reflected in revised engineering and code design, will require coordinated actions over time. Education and outreach activities are needed to ensure stakeholders, including community officials, engineers and architects, building owners and the public at large, have the necessary tools to make effective decisions about the value of enhanced performance by designing to IO performance objectives.

Managing the Distribution of Costs and Benefits. One of the core challenges in constructing for enhanced building performance for both new and existing buildings is that owners and developers who invest in IO performance measures may not be the primary beneficiaries of the investment. Research is needed to help clarify costs and benefits and to support development of innovative and feasible adoption mechanisms, such as financial incentives to offset investment costs, that can help balance costs and benefits for stakeholders including occupants, building owners and communities.

Influencing Private Owners. While the performance of individual buildings during a hazard event cumulatively affects the ability of a community to respond to and recover from the event, the majority of buildings are privately owned. Research is needed to identify the considerations associated with how private owners may be influenced or incentivized to participate in improving the performance of their buildings.

Influencing Public Sector. Buildings that are owned by local, state, or federal agencies (hospitals, nursing homes, housing, etc.) may affect community recovery, especially in economically disadvantaged regions. Public buildings may not be subject to local codes; thus, research is needed to identify appropriate implementation and adoption mechanisms for the public sector.

Protecting Vulnerable Populations. Vulnerable populations are more likely to live in older structures and often in more hazard-prone areas such as flood plains. It is important that adoption measures ensure all populations, including those who are socioeconomically disadvantaged, the elderly, and those requiring medical or caregiving attention, have opportunities to benefit from enhanced building performance and hazard resilience.

Addressing Liability for Building Performance. A building's systems may not perform as anticipated during a hazard event. While in some circumstances this may be due to error in design, construction, or maintenance of the structure, building system performance can be affected by factors beyond the control of the designer. For example, performance might be impacted by an extreme hazard level that is not considered in the design, the availability of infrastructure services, or other factors outside of the building envelope and beyond building code requirements. Additional research and stakeholder input is needed to address legal issues surrounding liability for the actual hazard performance of buildings and the influence these considerations will have on IO performance objective adoption.

Coordinating Interdisciplinary Collaboration. Due to the interdisciplinary nature of developing and implementing actions and measures for IO performance objectives, collaboration is needed across the array of stakeholders that have an interest in enhanced building performance. This includes collaboration across

disciplines, professions, and across public and private sectors within a community. This collaborative approach is often challenging due to the traditional roles and responsibilities of individuals involved in building design. Harnessing the diverse set of relevant expertise is essential to ensure IO performance objectives are adopted in an effective, successful manner.

Garnering Public Support. Stakeholder support is critical to the success of achieving IO performance. Eliciting buy-in and support across individuals, public and private sectors, and communities, is essential to garnering community trust, participation, and influence in developing IO building performance initiatives. By collaborating with existing community networks and leveraging the role of community leaders, local knowledge, skills, resources, and priorities can more effectively be integrated to achieve IO goals.

Conclusion

This paper articulates crosscutting issues as well as key policy challenges to support the development and implementation of IO performance objectives. The crosscutting issues are critical steps toward IO development as they cover multiple technical fields and also consider the interaction between building design, community resilience, social and economic impacts, and implementation activities. The key implementation challenges in successful adoption of IO performance objectives includes complex social, economic, and policy challenges. The challenge of achieving IO performance is just as much a social and economic matter as it is a technical one. The adoption of IO performance objectives will require holistic consideration of the impact of IO building performance on private owners, public sectors, and vulnerable populations. In addition, garnering support for, and shifting public expectations to, IO performance is important for successful implementation of IO objectives.

The diverse research needs and challenges discussed in the paper demand multidisciplinary perspectives and engagement from all levels of society. They will require reallocation of existing effort, time, resources, and financial investment. Moreover, substantial changes would be required for education, training, and practice within the engineering, architectural, and building professions. The involvement and enthusiasm of professional societies and other key stakeholders would be necessary to produce change within standards developing organizations and in building codes. While these activities are necessary for achieving IO performance objectives, additional research and implementation activities concerning the performance of infrastructure and interaction of infrastructure with the functionality of IO buildings would be needed to improve the resilience of buildings to the benefit of the public.

Acknowledgements

This project was completed with assistance from personnel at the Science and Technology Policy Institute (STPI) of the Institute for Defense Analyses (IDA), located in Alexandria, VA. The authors also gratefully acknowledge contributions from a number of groups: the steering committee members Mary Comerio, Gregory Deierlein, Susan Dowty, John Gillengerton, James Harris, William Hirano, Laurie Johnson, Timothy Reinhold, and James Rossberg; workshop participants; and reviewers of the NIST report.

References

FEMA, 2017, *Hazus Estimated Annualized Earthquake Losses for the United States*, FEMA P-366, Federal Emergency Management Agency, Washington, DC, U.S.A., 75 pages.

Mooney, C., Dennis, B., 2018, “Extreme hurricanes and wildfires made 2017 the most costly U.S. disaster year on record,” Washington Post, Energy and Environment, Washington, DC, U.S.A.,

https://www.washingtonpost.com/news/energy-environment/wp/2018/01/08/hurricanes-wildfires-made-2017-the-most-costly-u-s-disaster-year-on-record/?utm_term=.89276e56808a, Accessed January, 2018.

Munich Re, 2018, "Natural catastrophe review: Series of hurricanes makes 2017 year of highest insured losses ever," Munich, Germany, https://www.munichre.com/en/media-relations/publications/press-releases/2018/2018-01-04-press-release/index.html, Accessed January, 2018.

NOAA National Centers for Environmental Information (NCEI) U.S. Billion-Dollar Weather and Climate Disasters, 2018, Washington, DC, https://www.ncdc.noaa.gov/billions/.

Sattar, S., McAllister, T., Johnson, K., Clavin, C., Segura, C., McCabe, S., Fung, J., Abrahams, L., Sylak-Glassman, E., Levitan, M., Harrison, K., Harris, J., 2018, *Research Needs to Support Immediate Occupancy Building Performance Following Natural Hazards*, NIST SP-1224, National Institute of Standards and Technology, Gaithersburg, MD, U.S.A., 80 pages.

U.S. Senate; S. Rep. No. 114-239, at Disaster Resilient Buildings, 2016, Retrieved from GPO's Federal Digital System: https://www.gpo.gov/fdsys/pkg/CRPT-114srpt239/html/CRPT-114srpt239.htm.

Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation

Jordan S. Weaver, Meir Kreitman, Jarred C. Heigel and M. Alkan Donmez

Abstract Laser-based additive manufacturing of metals relies on many micro-sized welds to build a part. A simplified, well-studied case of this process is a single scan of the laser across a single layer of powder. However, there is a lack of mechanical property measurements of the tracks produced in such experiments. Nanoindentation measurements on laser track cross sections of nickel superalloy 625 reveal hardness differences between the track melt pool and base material as well as variations with laser scan speed. There is a change from ≈5.5 GPa in the track melt pool to ≈4.8 GPa in the base material for laser settings of 195 W and 800 mm s^{-1}. In comparison, the increase in hardness in the melt pool is not observed for settings of 195 W and 200 mm s^{-1}. It is believed that the difference in thermal histories supported by thermographic measurements causes a difference in the dislocation density in the melt pool. This results in a difference in hardness between the two tracks. The effects of the local crystal orientation, dendritic spacing, and residual stress are considered in the interpretation of results.

Keywords Additive manufacturing · Hardness · Residual stress · Microstructure · Electron backscatter diffraction

Introduction

Laser-based additive manufacturing of metals relies on the process of a laser to melt feedstock material to build a component layer by layer. This process can be broken down to the fundamental step of a single scan of the laser across a bare plate or a single layer of powder. These rather simple experiments have proven highly useful for developing models that can predict temperature, residual stress, melt pool geometry, and microstructure (e.g., [1–8]). Less emphasis has been placed on predicting mechanical properties (e.g., hardness) of single scan laser tracks. This may be in part due to the difficulties associated with testing and interpreting mechanical performance over

Weaver, Jordan; Kreitman, Meir; Heigel, Jarred; Donmez, M.
"Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation."
Paper presented at TMS 2019, San Antonio, TX, United States. March 10, 2019 - March 14, 2019.

micrometer length scales. While understanding the mechanical performance of built components is the goal, there is still a lot to gain from understanding the mechanical properties inside single scan laser tracks. In such experiments, it is possible to clearly link the mechanical response to the laser power and scan speed without the complications of the compound thermal history that exists in a component.

Materials and Methods

Nickel superalloy 625 (IN625) was chosen because it is widely used in additive manufacturing and industrial applications. IN625 plate of approximately 3.2 mm thick was polished to 400 grit and annealed at 870 °C for 1 h in vacuum. A single layer of EOS[1] IN625 powder was hand spread with an approximate layer thickness of 36 μm prior to exposing to the laser. Single scans on the single layer of powder were made following procedures given in Ref. [8] using a commercial EOS M270D[1] laser bed powder fusion machine. Each single scan was sufficiently spaced to reduce the influence on neighboring scans and sufficiently long to reach a steady state. Two different combinations of laser power and scan speed were investigated: 195 W at 200 mm s^{-1} and 195 W at 800 mm s^{-1}. The estimated laser spot size for these experiments and the single scan laser tracks in Ref. [8] is 100 μm. This is different than the estimated spot size of 188 μm for experiments in Ref. [3]. The difference in spot size and the addition of a layer of powder should be kept in mind when comparing the melt pool geometries from this study to Ref. [3].

Single scan laser tracks were cross sectioned as shown in Fig. 1a, b. The cross sections were mounted and metallographically prepared using a final vibratory polish with 0.02 μm colloidal silica. Nanoindentation was performed using an MTS (Keysight) Nanoindenter XP[1] on cross sections in the melt pool regions and far from the melt pools as shown in Fig. 1b. Here, we use the term melt pool to describe the material that was melted by the laser and resolidified. A final indentation depth of 300 nm was chosen to reduce the spacing between indents to 10 μm allowing for at least several indents inside the melt pool. This produces residual indents on the order of a micrometer (see Fig. 1c, d) which are influenced by the local crystal structure. The local crystal structure at each indentation site was determined from electron backscatter diffraction (EBSD) using a JOEL JSM7100[1] field-emission scanning electron microscope (SEM) and Oxford[1] EBSD detector (Fig. 1c, d). The positions of indents with respect to the surface and melt pool geometry were determined from optical micrographs before and after etching using a Zeiss LSM 800[1] optical microscope. Etching with aqua regia was necessary to reveal the melt pool boundary to identify indents as either inside or outside the melt pool.

[1]Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Weaver, Jordan; Kreitman, Meir; Heigel, Jarred; Donmez, M.
"Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation."
Paper presented at TMS 2019, San Antonio, TX, United States. March 10, 2019 - March 14, 2019.

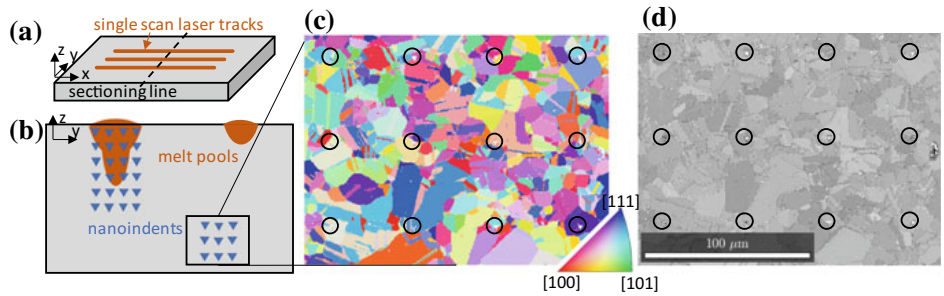


Fig. 1 **a** Schematic of single scan laser tracks on a single layer of powder, **b** schematic of single track laser scan cross sections with nanoindents. The melt pool size/shape varies depending on the laser power and scan speed. Indent size is exaggerated. **c** EBSD inverse pole figure map of the base material far from the laser tracks containing circled indents to 300 nm displacement, **d** corresponding band contrast image. The scale bar is the same for (**c**) and (**d**)

Nanoindentation was performed with a diamond Berkovich tip to a final displacement of 300 nm. The strain rate, which is the loading rate divided by the load, was held constant at 0.05 s^{-1}. The continuous stiffness method (CSM), which superimposes a small oscillatory loading signal on the monotonic loading signal, was employed with a displacement amplitude of 2 nm and frequency of 45 Hz. The CSM allows for the determination of the unloading stiffness, S, throughout the test from many small unloads [9]. Unloading is necessary to determine the contact area, A, effective modulus, E_{eff}, and hardness, H, in accordance with the Oliver–Pharr analysis [10]. The hardness in Eq. (1) is simply the load divided by the cross-sectional area. The area function is determined from tests of fused quartz [10]. All eight area function coefficients were used with an emphasis on data at shallower depths <1000 nm. The effective modulus, Eq. (2), depends on the stiffness, contact area, and a correction factor β. A value of 1.034 was used in the area function calibration and analysis [10]. The effective modulus, Eq. (3), is the elastic response of the indenter tip and sample designated with subscripts i and s, respectively, which is dependent on the moduli, E, and Poisson ratios, v. A modulus and Poisson's ratio of 1140 GPa and 0.07, respectively, were used for the diamond tip. A sample Poisson's ratio of 0.31 was used to determine the sample modulus. This value is the Voigt–Reuss–Hill Poisson's ratio [11] based on available elastic constants of nickel superalloy 625 [12]. Note that the sample modulus, E_s, is not the single crystal Young's modulus. A more rigorous account of elastic anisotropy of cubic crystals during indentation should be followed if this is desired (see Refs. [13, 14]).

$$H = \frac{P}{A} \tag{1}$$

$$E_{eff} = \frac{\sqrt{\pi}S}{2\beta\sqrt{A}} \tag{2}$$

$$\frac{1}{E_{eff}} = \frac{1-v_s^2}{E_s} + \frac{1-v_i^2}{E_i} \tag{3}$$

Weaver, Jordan; Kreitman, Meir; Heigel, Jarred; Donmez, M.
"Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation."
Paper presented at TMS 2019, San Antonio, TX, United States. March 10, 2019 - March 14, 2019.

Indentation size effects, where the hardness is higher at shallower depths is commonly reported in metals and alloys [15]. A model that can be used to estimate the intrinsic hardness or hardness at infinite depth is the Nix–Gao model using Eq. (4) [15]. The intrinsic hardness, H_0, and length scale parameter, h^*, are determined from a regression of the hardness, H, and displacement, h, data.

$$\frac{H}{H_0} = \sqrt{1 - \frac{h^*}{h}} \tag{4}$$

The apparent indentation size effect at very shallow depths <200 nm is sensitive to the sharpness of the tip, sample preparation, etc. [16]. The tests on the laser track cross sections were limited to 300 nm to reduce the spacing between indents and to fit many into the melt pool. This leaves only a small portion of data (200–300 nm displacement) to determine the intrinsic hardness and length scale parameter. Tests on the annealed plate up to 1000 nm displacement were also analyzed. The mean values using data from 200 to 300 nm and 200 to 1000 nm were similar while the variance was significantly higher for 200–300 nm. This should be kept in mind when interpreting the Nix–Gao parameters from a limited range of indentation depth.

Results and Discussion

Figure 2 shows the shape and grain structure of the laser track cross sections from EBSD and optical micrographs. The most obvious difference between the two tracks is in the size of the melt pool with clear keyholing occurring for slower scan speed. The melt pool boundaries on the EBSD maps are approximations based on the formation of the elongated grain structure in the melt pool. The actual melt pool boundary is determined from the optical images of etched samples. Nanoindents can be seen on the band contrast maps (Fig. 2b, e) as a uniform grid of black spots with 10 μm spacing. They are also faintly present in the optical images which allowed for categorizing indents as either inside the melt, outside the melt, or on the border of the two regions.

Figure 3 shows the nanoindentation trends for one laser track, 195 W at 800 mm s^{-1}. The average hardness, Fig. 3a, and sample modulus, Fig. 3b, for each test were determined using data between 275 nm and 300 nm displacement. The Nix–Gao model values, Fig. 3c, d, were determined by a regression to data between 200 and 300 nm displacement. The x-axis in these plots is the perpendicular distance each indent is from the original surface of the plate. Some indents can have a negative position if they landed in the solidified powder layer above the plate's original surface. In addition, each data point or indent is categorized as either inside the melt pool, outside the melt pool, or on the border. Indents were considered on the border if they fell directly on the melt pool boundary or fell relatively close to the boundary (≈3 μm). There is a clear increase in hardness and slightly lower modulus in the melt pool (Fig. 3a) compared to outside the melt pool. This is likely due to a combination

Weaver, Jordan; Kreitman, Meir; Heigel, Jarred; Donmez, M.
"Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation."
Paper presented at TMS 2019, San Antonio, TX, United States. March 10, 2019 - March 14, 2019.

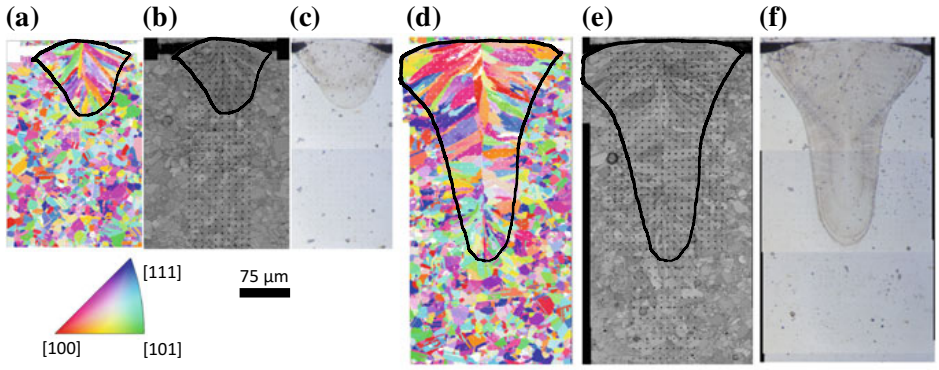


Fig. 2 **a** EBSD inverse pole figure, **b** EBSD band contrast, and **c** the corresponding optical micrograph of laser track cross sections after nanoindentation for 195 W at 800 mm s^{-1}. **d** EBSD inverse pole figure, **e** EBSD band contrast, and **f** the corresponding optical micrograph of laser track cross sections after nanoindentation for 195 W at 200 mm s^{-1}

of increased dislocation density, the dendritic microstructure, and residual stresses compared to the annealed plate where these do not exist (dendrites, residual stress) or are minimized (dislocation density). The temperature gradient that occurs in the melt pool during the rapid solidification produces internal stresses. These stresses during the solidification process are likely high enough to generate dislocations through local plastic deformation and/or crystal orientation gradients. An increase in dislocation density will increase the measured hardness (i.e., Taylor hardening law). The intrinsic hardness, Fig. 3c is also higher in the melt pool for the same reasons and the length scale parameter is reduced (i.e., the indentation size effect is reduced). Other scenarios where the intrinsic hardness increases and the length scale parameter reduces are in cold-worked [15] and radiation-damaged materials [17]. In both these scenarios, there is an increase in the dislocation density in the material.

The residual internal stresses in the material after the material solidifies and cools can also affect the hardness as well as the measured modulus [18]. Generally, a compressive residual stress in the indentation plane will increase the measured hardness and a tensile stress will reduce the hardness [18]. Changes in the measured sample modulus are a good indicator of this type of residual stress and severity, with compressive residual stresses increasing the effective modulus and tensile residual stress decreasing the effective modulus [18]. However, Ref. [18] found that the measured changes in modulus and hardness with residual stress went away when the contact area was determined from images of residual indents rather than the unloading stiffness. The conclusions are likely dependent on the tip geometry, indentation depth, and material making it difficult to directly apply them to this study. Another solution to this issue that does not require measuring the residual indent area is to use the known modulus to correct the data [19]. Here, we leave the measurements as is and caution that whenever the modulus values deviate from the base material without a physical reason such as the case of residual stresses, there may be errors in the hardness associated with errors in the contact area. This issue requires further investigation.

Weaver, Jordan; Kreitman, Meir; Heigel, Jarred; Donmez, M.
"Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation."
Paper presented at TMS 2019, San Antonio, TX, United States. March 10, 2019 - March 14, 2019.

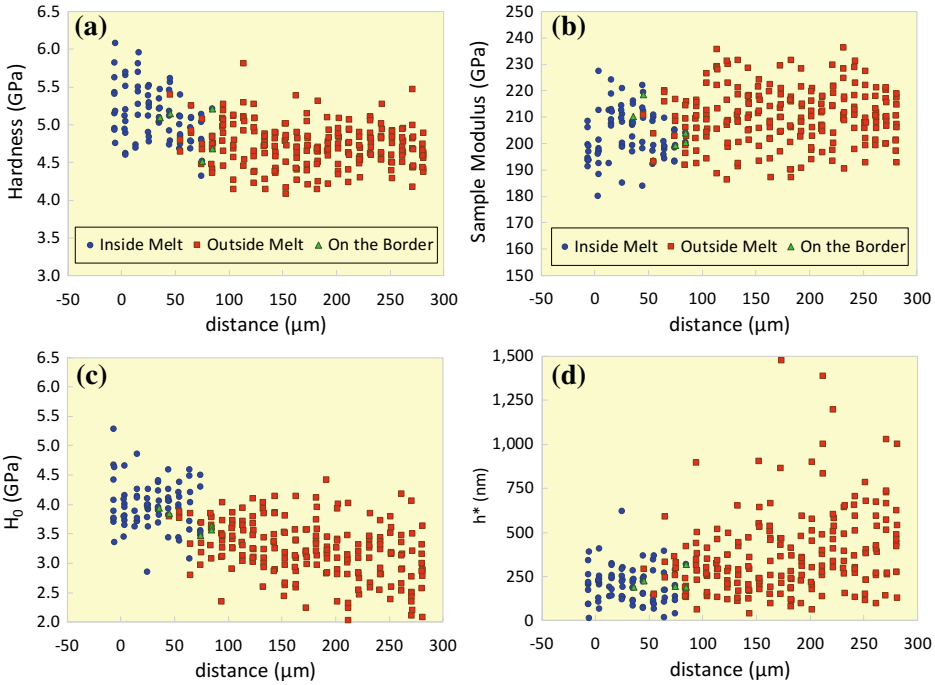


Fig. 3 Indentation property versus the distance of each indent from the laser track surface for 195 W at 800 mm s^{-1}: **a** hardness at 275–300 nm displacement, **b** modulus at 275–300 nm displacement, **c** intrinsic hardness, H_0, determined from 200 to 300 nm displacement and **d** length scale parameter, h^*, determined from 200 to 300 nm displacement. The blue circles are indents that fell inside the melt pool, green triangles were on the border, and the red squares fell outside the melt pool

Figure 4 shows the indentation trends for laser track 195 W at 800 mm s^{-1} reduced to tests inside similar grain orientations. Grains were defined from EBSD data based on a misorientation angle <5°. Grains with a crystal normal within 10° of the (2 2 11) direction were isolated. This crystal plane was chosen so that several grains inside and outside the melt pool could be compared. Some of the indents considered are very close to grain boundaries. A stricter criterion would be to only consider indents that are approximately three times the residual indent diameter away from any grain boundaries; however, this would eliminate most of the tests. Rather we consider tests if they did not fall directly on any grain boundaries and meet the orientation requirement. The reduced data based on similar grain orientations shows the same trends as the grid of indents which does not consider grain orientation. This means the grain orientation does not have a significant effect on the nanohardness trends. We also note that even inside the same grain in the melt pool (e.g., indent numbers 4, 18, and 21), the hardness and modulus measurements vary. For these reasons, the arrays of indents are sufficient for comparison of different laser tracks. This may not be the case for crystals with greater plastic anisotropy (e.g., hexagonal crystals) or indentation with tip geometries that produce less plastic deformation (e.g., spherical tips).

Weaver, Jordan; Kreitman, Meir; Heigel, Jarred; Donmez, M.
"Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation."
Paper presented at TMS 2019, San Antonio, TX, United States. March 10, 2019 - March 14, 2019.

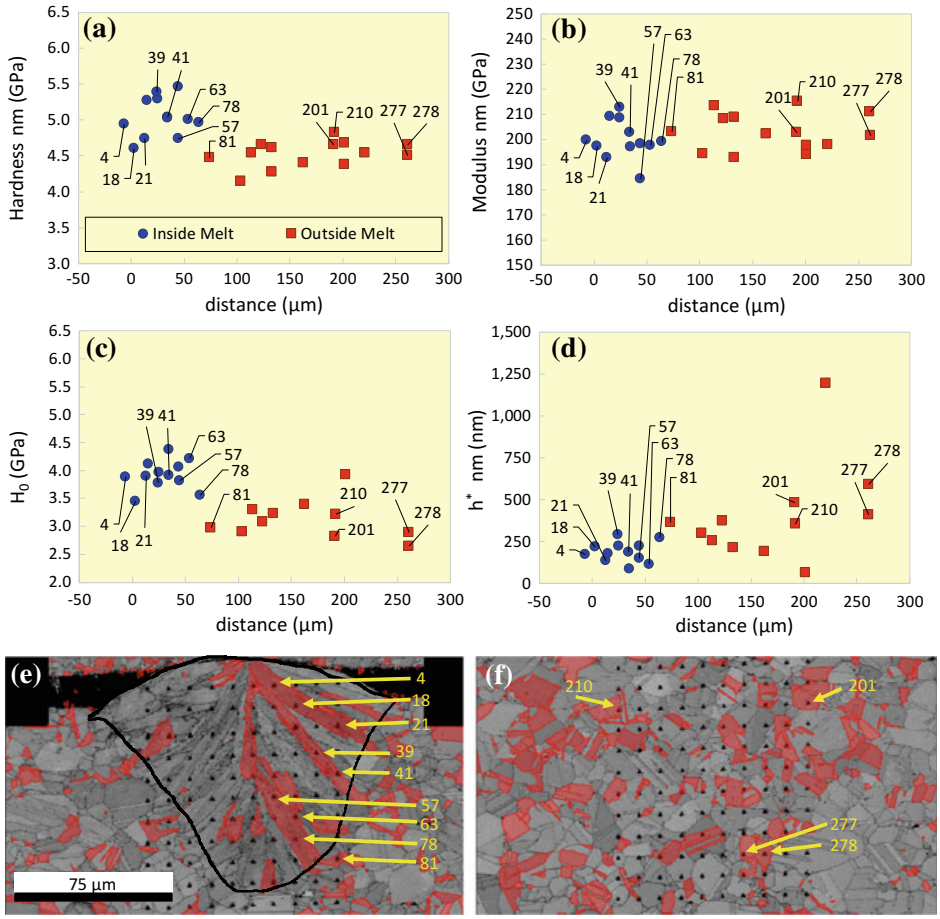


Fig. 4 Indentation property versus the distance of indents inside grains with a crystal plane <10° from (2 2 11) for the laser track 195 W at 800 mm s^{-1}: **a** hardness, **b** modulus, **c** intrinsic hardness, H_0, and **d** length scale parameter, h^*. The blue circles are indents that fall inside the melt pool, and the red squares fall outside the melt pool. (**e**, **f**) EBSD band contrast images with grains shaded red that meet orientation criterion. A selection of indents is labeled with their respective number. Indents 4, 18, and 21 are inside one grain. Indents 39 and 41 are inside one grain. Indents 57, 63, 78, and 81 are also inside one grain

Figure 5 shows a direct comparison of the hardness and modulus for the two laser tracks at a power of 195 W and scan speeds of 800 and 200 mm s^{-1}. The dashed lines in the plots are based on one standard deviation above and below the mean of 95 indents on the annealed plate far from the tracks. The increase in hardness on the track with scan speed of 800 mm s^{-1} is not seen in the track with scan speed of 200 mm s^{-1}. To understand this, we consider the difference in the thermal history between the two tracks. Radiant temperature measurements along tracks and estimated cooling rates were made using the procedures in Ref. [8] which show that the temperature gradients, change in temperature over distance, trailing the melt pool are similar between the two tracks. The cooling rate, change in temperature

Weaver, Jordan; Kreitman, Meir; Heigel, Jarred; Donmez, M.
"Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation."
Paper presented at TMS 2019, San Antonio, TX, United States. March 10, 2019 - March 14, 2019.

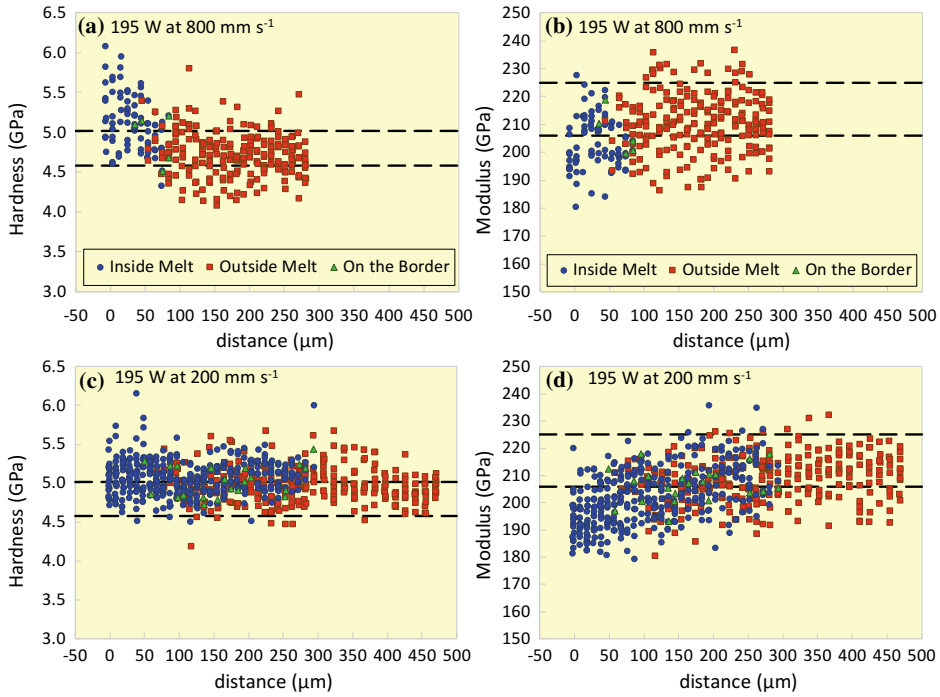


Fig. 5 Comparison of indentation hardness (**a**, **c**) and sample modulus (**b**, **d**) for laser tracks 195 W. **a**, **b** 800 mm s^{-1}, **c**, **d** 200 mm s^{-1}. The dotted lines are ± one standard deviation of the mean for 95 indents on the annealed plate far from the tracks. **a**, **b** contain 295 indents and **c**, **d** contain 690 indents

over time, is roughly the temperature gradient multiplied by the laser scan speed. This means that the estimated cooling rate scales with the laser scan speed; the cooling rate is approximately four times greater for the 800 mm s^{-1} track compared to the 200 mm s^{-1} track. A difference in cooling rate will produce a difference in the dendritic spacing in the melt pool which can be estimated from phase field simulations as 0.2 μm and 0.43 μm for the 800 mm s^{-1} and 200 mm s^{-1} scan speeds, respectively [3]. The indents are on the order of 1 micrometer such that in both cases it is likely probing multiple dendrites. We reason that the difference in dendritic spacing is not a significant factor in the comparison of hardness between the two tracks. It should be noted that the radiant temperature measurements are surface measurements and do not measure the temperature gradient or cooling rate along the depth of the melt pool. Based on the radiant temperature measurements and the size of the melt pool cross sections, we reason that the thermal histories are sufficiently different to cause differences in the dislocation density. A greater cooling rate, possibly steeper temperature gradient along the depth of the melt pool, and subsequent greater dislocation density inside the 800 mm s^{-1} track would explain why the hardness is higher in the 800 mm s^{-1} track compared to the 200 mm s^{-1} track.

Weaver, Jordan; Kreitman, Meir; Heigel, Jarred; Donmez, M.
"Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation."
Paper presented at TMS 2019, San Antonio, TX, United States. March 10, 2019 - March 14, 2019.

The modulus is reduced for both tracks in the melt pool to a similar level. Any error in hardness due to this effect, as discussed earlier, likely does not have a significant effect in the comparison between tracks. The uncertainty in the modulus and hardness values is estimated based on Ref. [20] which found an average uncertainty (one standard deviation) among individual participants in a round robin study to be 4% of the average hardness and 5% of the average modulus. For comparison, one standard deviation of 95 measurements on annealed plate material far from the laser tracks was 0.22 GPa or 4.7% of the average hardness and 10 GPa or 4.4% of the average modulus.

Summary

Two different single scan laser tracks on a single layer of nickel superalloy 625 powder were cross sectioned and characterized with Berkovich nanoindentation. EBSD was used to isolate grain orientation effects and optical microscopy was used to determine position of indents relative to the surface and melt pool boundary. There are several findings from these experiments which are as follows:

1. The hardness inside the melt pool for 195 W, 800 mm s^{-1} is higher than the annealed plate likely due to an increase in dislocation density in the melt pool because of the rapid cooling and temperature gradients during the solidification process.
2. Isolating indents based on similar grain orientations show the same trend across the melt pool as large arrays of indents. Crystal orientation effects on the indentation response are not a significant factor in interpreting this set of experiments.
3. The comparison of the hardness and modulus for laser tracks of 195 W at 800 and 200 mm s^{-1} reveals that the hardness is higher inside the melt pool for a scan speed of 800 mm s^{-1}. The increased cooling rate and possibly steeper temperature gradient along the melt pool depth in the 800 mm s^{-1} track increase the dislocation density and subsequent hardness.
4. The measured modulus is reduced in the laser tracks likely due to residual stresses. This might cause the hardness to be underestimated due to an error in the determination of the contact area. The reduction in modulus is similar for both tracks and does not affect the comparison between the two different tracks.

Acknowledgements We wish to thank Will Osborn and Maureen Williams of the Materials Measurement Laboratory at NIST for their help with EBSD. We are very appreciative of Mark Stoudt of the Materials Measurement Laboratory at NIST for etching samples and discussing the many aspects of additive nickel superalloy 625. We are also grateful of Lyle Levine and Will Osborn of the Materials Measurement Laboratory at NIST for bringing to our attention the need to do these experiments. Meir Kreitman wishes to acknowledge support from the University of Maryland and the NIST Summer Undergraduate Research Fellowship (SURF) program.

Weaver, Jordan; Kreitman, Meir; Heigel, Jarred; Donmez, M.
"Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation."
Paper presented at TMS 2019, San Antonio, TX, United States. March 10, 2019 - March 14, 2019.

References

1. Hussein A, Hao L, Yan C, Everson R (2013) Finite element simulation of the temperature and stress fields in single layers built without-support in selective laser melting. Mater Des (1980–2015) 52:638–647
2. Keller T, Lindwall G, Ghosh S, Ma L, Lane BM, Zhang F, Kattner UR, Lass EA, Heigel JC, Idell Y, Williams ME, Allen AJ, Guyer JE, Levine LE (2017) Application of finite element, phase-field, and CALPHAD-based methods to additive manufacturing of Ni-based superalloys. Acta Mater 139:244–253
3. Ghosh S, Ma L, Levine LE, Ricker RE, Stoudt MR, Heigel JC, Guyer JE (2018) Single-track melt-pool measurements and microstructures in Inconel 625. JOM 70(6):1011–1016
4. Montgomery C, Beuth J, Sheridan L, Klingbeil N (2015) Process mapping of Inconel 625 in laser powder bed additive manufacturing. In: Solid freeform fabrication symposium, pp 1195–1204
5. Heigel JC, Lane BM (2018) Measurement of the melt pool length during single scan tracks in a commercial laser powder bed fusion process. J Manuf Sci Eng 140(5):051012–051012-7
6. Akram J, Chalavadi P, Pal D, Stucker B (2018) Understanding grain evolution in additive manufacturing through modeling. Addit Manuf 21:255–268
7. Ma L, Fong J, Lane B, Moylan S, Filliben J, Heckert A, Levine L (2015) Using design of experiments in finite element modeling to identify critical variables for laser powder bed fusion. In: International solid freeform fabrication symposium, laboratory for freeform fabrication and the University of Texas Austin, TX, USA, pp 219–228
8. Heigel JC, Lane BM (2017) The effect of powder on cooling rate and melt pool length measurements using in situ thermographic techniques. In: Solid freeform fabrication symposium
9. Hay J, Agee P, Herbert E (2010) Continuous stiffness measurement during instrumented indentation testing. Exp Tech 34(3):86–94
10. Oliver WC, Pharr GM (2004) Measurement of hardness and elastic modulus by instrumented indentation: advances in understanding and refinements to methodology. J Mater Res 19(1):3–20
11. Hill R (1952) The elastic behaviour of a crystalline aggregate. Proc Phys Soc Sect A 65(5):349
12. Wang Z, Stoica AD, Ma D, Beese AM (2016) Diffraction and single-crystal elastic constants of Inconel 625 at room and elevated temperatures determined by neutron diffraction. Mater Sci Eng A 674:406–412
13. Vlassak JJ, Nix WD (1994) Measuring the elastic properties of anisotropic materials by means of indentation experiments. J Mech Phys Solids 42(8):1223–1245
14. Vlassak JJ, Nix WD (1993) Indentation modulus of elastically anisotropic half-spaces. Philos Mag A 67(5):1045–1056
15. Nix WD, Gao HJ (1998) Indentation size effects in crystalline materials: a law for strain gradient plasticity. J Mech Phys Solids 46(3):411–425
16. Pharr GM, Herbert EG, Gao YF (2010) The indentation size effect: a critical examination of experimental observations and mechanistic interpretations. Annu Rev Mater Res 40:271–292
17. Hosemann P, Shin C, Kiener D (2015) Small scale mechanical testing of irradiated materials. J Mater Res 30(9):1231–1245
18. Tsui TY, Oliver WC, Pharr GM (2011) Influences of stress on the measurement of mechanical properties using nanoindentation: Part I. Experimental studies in an aluminum alloy. J Mater Res 11(3):752–759
19. Hou X, Jennett N (2017) A method to separate and quantify the effects of indentation size, residual stress and plastic damage when mapping properties using instrumented indentation. J Phys D Appl Phys 50(45):455304
20. Read DT, Keller RR, Barbosa N, Geiss R (2007) Nanoindentation round robin on thin film copper on silicon. Metall Mat Trans A 38(13):2242–2248

Weaver, Jordan; Kreitman, Meir; Heigel, Jarred; Donmez, M.
"Mechanical Property Characterization of Single Scan Laser Tracks of Nickel Superalloy 625 by Nanoindentation."
Paper presented at TMS 2019, San Antonio, TX, United States. March 10, 2019 - March 14, 2019.

Proceedings of the ASME 2019 International Manufacturing Science and Engineering Conference
MSEC2019
June 10 – June 14, 2019, Erie, Pennsylvania, USA

MSEC2019-2748

TESTING OF THE MTCONNECT – OPC-UA COMPANION SPECIFICATION

Ryan Fisher
Department of Aerospace Engineering,
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061 U.S.A.

Guodong Shao
Systems Integration Division
National Institute of Standards and Technology
Gaithersburg, Maryland 20899 U.S.A.

KEYWORDS
Companion Specification; Interoperability; MTConnect; OPC-UA; Testing.

ABSTRACT
Smart Manufacturing (SM) is the future of the manufacturing industry. Seamless, accurate, and fast connection and communications among devices are critical for SM. By leveraging information technologies, devices can dynamically communicate with each other to increase factory production, while decreasing engineering costs. MTConnect and Open Platform Communications - Unified Architecture (OPC-UA) standards facilitate such communication. MTConnect is a manufacturing interoperability standard that provides a semantic vocabulary for manufacturing equipment to provide structured contextualized data with no proprietary format. The OPC-UA is a platform-independent standard through which various systems and devices can communicate by sending messages between clients and servers over various networks. OPC-UA enables syntactic interoperability between clients and servers. The MTConnect - OPC-UA Companion Specification integrates the two standards to provide manufacturers more efficient and powerful interoperability capabilities. In this paper, we report the test of version 1.02 of this companion specification. This specification sets a standard means of communication between MTConnect devices and OPC-UA Clients/Servers based on Extensible Markup Language (XML) structures. To test the standard, the following components have been developed: an OPC-UA Server, an OPC-UA Client, a probe that translates data structures in MTConnect XML format to MTConnect OPC-UA Companion XML format that can be recognized by the server, a MTConnect XML data parser, and a MTConnect device simulator. The activities of the standard testing include passing varying data structures and objects through the server and confirming the information is received accurately by the client. The findings of the standard testing will be provided to the standard developing organizations for improving the future versions of the standard.

INTRODUCTION
By increasing the availability of information, the processes of product design, manufacturing, and quality control can all be improved. Now, more than ever, there is a demand for larger amounts of real-time information on manufacturing floors of every scale. The dynamic information enables manufacturers to have better awareness and control over their machines and processes. The information is generated and communicated through the connection and integration of machines with other devices, systems, and applications (e.g., controllers, simulators, or Graphical User Interface (GUI)).

MTConnect [1] and the Open Platform Communications (OPC) are two standards that help achieve these goals in the manufacturing industry. MTConnect provides semantic vocabularies for manufacturing capabilities including, but not limited to, device monitoring, automation, and process analytics. The OPC Foundation has created the OPC-Unified Architecture (UA) [2] standard that supports object-oriented implementations that can handle information such as alarms and events, commands, real-time data, and complex data in Extensible Markup Language (XML). MTConnect and the OPC Foundation have partnered to develop the MTConnect - OPC-UA Companion Specification [3] to set a standard means of communication (a gateway) between MTConnect devices, servers, or agents, and OPC-UA servers and clients. Testing the MTConnect - OPC-UA Companion Specification could demonstrate the usability and feasibility of the standard for the manufacturing industry while identifying issues in the current version of the specification can provide feedback to the two standard developing organizations (SDOs), i.e., MTConnect and OPC Foundation, for improvements in future versions. Previous research conducted on manufacturing data integration claims

Fisher, Ryan; Shao, Guodong.
"TESTING OF THE MTCONNECT – OPC-UA COMPANION SPECIFICATION."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, Erie, PA, United States. June 10, 2019 - June 14, 2019.

that the integration of many devices with the companion standard in a factory-wide network is easier compared to small scale implementations; small scales were more efficient with the sole use of MTConnect [4]. The testing in this study may confirm such claims, however this is not a formal requirements-driven verification and validation (V&V) activity. It is a preliminary specification conformance testing by comparing the outputs of MTConect machine and the OPC-UA client. Conducting such a test requires a cross-platform testing environment that allows a variety of test cases to be performed. The cross-platform capability is essential to the OPC-UA components. The test cases defined will represent a variety of data types and data items to highlight key aspects of the specification including Data Access, Historical Access, and Event and Alarm Notifiers.

This paper describes the development of the testing environment and the testing components including an OPC-UA server, an OPC-UA client, an MTConnect probe, an MTConnect data parser, a MTConnect device simulator, and test cases. These components can also be reused for testing future version of companion specification. The test cases are typically applicable to all versions, as the MTConnect data model will not change.
The organization of this paper is as follows. The relevant standards, i.e., MTConnect, OPC-UA, and the MTConnect - OPC-UA companion specification, are introduced first. The next Section explains the methodology of testing the companion specification and discusses how the testing environment and its components were established and challenges that were encountered during the testing. The three key features of OPC-UA are also discussed. Then, the following Section identifies test cases and provides a justification for such test cases. After that, the next Section presents and discusses the testing results of the companion specification. The final Section provides a conclusion and discusses the future work.

RELEVEANT STANDARDS

MTConnect and OPC-UA standards are two interoperability standards that facilitate communication among smart devices in manufacturing domain. Integrating the two standards provides manufacturing companies powerful interoperability capabilities. Each of the standards is briefly discussed as follows.

The MTConnect standard provides a semantic vocabulary for manufacturing equipment to generate structured, contextualized data with no proprietary format. With uniform data, users can focus on manufacturing applications rather than translation. MTConnect data sources include production equipment, sensor packages, and other factory floor hardware. MTConnect's Extensible Markup Language (XML) data format provides both human and machine-readable features. MTConnect is extensible and can be integrated with other standards by design, which facilitate the integration with OPC-UA [1].

OPC-UA is a platform-independent service-oriented architecture that integrates all the functionality of the individual OPC Classic specifications into one extensible framework. It provides the equivalent functions of the original OPC Classic, while extending the object-oriented capabilities to complex and multi-level structures, but with more functionality: on-demand data access, data subscriptions, event notifiers, method executions, and server discovery. Through OPC-UA, various systems and devices can communicate by sending messages between clients and servers over various networks. OPC-UA enables syntactic interoperability between clients and servers. The communication is provided on a safe and secure network by 128-bit or 256-bit encryption levels, message signing, user/system auditing, and authentication [2]. OPC-UA uses a predefined semantic vocabulary represented in XML to provide a descriptive schema of how items are in general mapped to the object-oriented model.

The combination of MTConnect and OPC-UA into a companion specification provides a powerful means of connecting MTConnect-enabled machining tools to OPC-UA servers and clients. As well, the inverse can be stated; the companion specification gateway allows for integration between OPC-UA servers and MTConnect applications. We will connect MTConnect machine tools to OPC-UA servers and clients, as shown in Figure 1 from Companion Specification Version 1.02 [3]. A device such as a milling machine is linked directly to an MTConnect server. Using the MTConnect to OPC-UA Gateway, a connection can be established between the OPC-UA client and the MTConnect server, allowing for information transfer over the network.

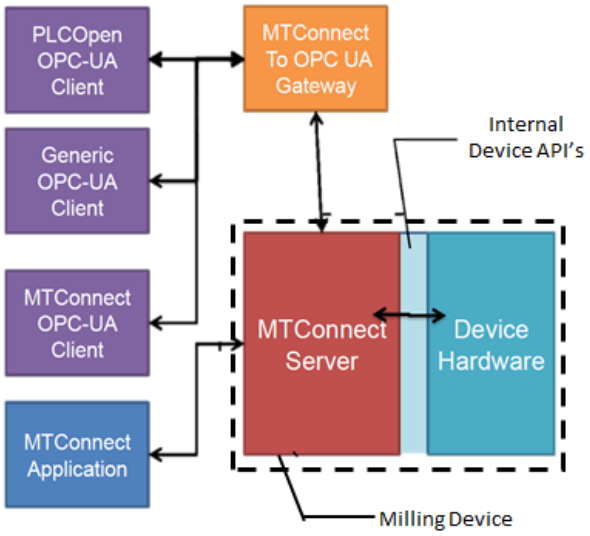


Figure 1. MTConnect to OPC-UA Gateway Implementation [3]

TESTING ENVIRONMENT IMPLEMENTATION

Fisher, Ryan; Shao, Guodong.
"TESTING OF THE MTCONNECT – OPC-UA COMPANION SPECIFICATION."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, Erie, PA, United States. June 10, 2019 - June 14, 2019.

To begin testing the MTConnect - OPC-UA Companion Specification, a testing environment including a server/client setup must be established. Figure 2 shows an information flow among the key components in the testing environment. MTConnect data is sent to the OPC-UA server via an XML parser and probe. The OPC-UA client retrieves the data from the server and sends the data to a GUI for visualization. The XML parser and probe also provides the data to the GUI for visualization and data comparison with the client output.

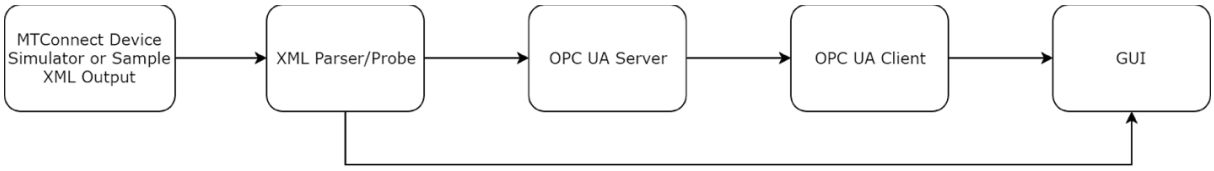


Figure 2. Companion Specification Implementation

MTConnect Device Simulator

There are several options for obtaining MTConnect data for a testing environment. For example, it can be real-time data from the MTConnect devices, data from online agent simulators, or data from software driven simulation agents. We have used the online agent simulator to retrieve device data. The online agent simulator is provided by the National Institute of Standard and Technology (NIST) and MemexOEE [5]. These sites provide the standard MTConnect XML files that contain a schema, a probe, real-time streams, and samples. The schema and probe have different XML structures compared to the stream file: the schema or probe file is required for the XML probe to generates nodes, and the stream is required for the parser to obtain data.

MTConnect Data Parser

Establishing MTConnect devices within an OPC-UA server requires the server to identify the structure of the device. For example, the server must know the component names such as Axes, and how many subcomponents each component contains, e.g., the Axes may contain Linear X Axis, Linear Z Axis, and Rotary C Axis. These objects have a parent-child relationship. Each component or subcomponent will have a child called "DataItem" at the end of its line of children components. Each of these components translates to an object in OPC-UA, while the DataItems translate to either variables or properties, all of which must be accounted for when generating nodes since each item becomes a node.

The data from the MTConnect device is in the form of an XML file, but it is being taken from the internet via a Uniform Resource Locator (URL). Using a URL to XML module [6], the MTConnect stream file is imported and filtered through the ElementTree XML API [7]. The ElementTree module allows for an XML file to be imported and its children can be analyzed from the root down, by exploring either the attributes, the tags, or the values. The DataItem tags are filtered via their attributes to locate the proper value. All values in XML are strings, which provides an ease of exportation to text files with a naming convention that uses a combination of component names and attributes to properly label data. The timestamp associated with each value is exported as a separate text file with identical component names and attributes but with a time extension added. It provides any other programs the ability to locate the data value along with its corresponding timestamp. For the purpose of visualization, timestamps are converted to a decimal version: 03:15 becomes 03.25.

MTConnect Data Probe

To search the XML MTConnect file for its components and DataItems, an XML probe is developed using Python and the ElementTree module. Using ElementTree, the probe generates an XML "companion file" to define a node structure for all devices in the MTConnect file, the XML companion file conforms to the OPC-UA "MTConnectModel.xml," found in the model compiler stack. The MTConnectModel.xml file represents basic MTConnect devices, components, and DataItems in the OPC-UA format. The .NETStandard server takes in only .uanodes files, which are binary file representations of the XML predefined nodes. The XML companion file needs to be converted to this binary format via a model compiler, which is provided by the OPC Foundation [8]. Conveniently, the .uanodes file, along with other files created by the model compiler, are automatically stored to the directory of the OPC-UA server. By simply setting a file path and executing the probe, the server can be started, and the nodes will automatically appear in the address space. A sample address space of the NIST Test Bed, obtained using UaExpert, is shown in Figure 3. Each node is automatically generated using either the MTConnect name or id attribute.

OPC-UA Server

The OPC-UA server in the testing environment is developed using the UA.NET Standard provided by the OPC Foundation [9]. There are a couple of options for the server development. We selected this option because of the code stacks are free and well-documented. In the package, QuickStart applications are provided for quickly generating a server using .NET in C#. For

Fisher, Ryan; Shao, Guodong.
"TESTING OF THE MTCONNECT – OPC-UA COMPANION SPECIFICATION."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, Erie, PA, United States. June 10, 2019 - June 14, 2019.

example, a boiler server conducts simulations of boilers, and the boiler objects in the address space are created using the OPC-UA model of object-typing and object instantiation. Once instantiation is completed, the simulation locates the nodes and establishes the values. Similarly, device nodes can be instantiated if the server reads in a file type of ".uanodes". Note that nodes can either be predefined or generated within the server configuration, however, the devices will have predefined nodes that are generated using the model compiler. Once the predefined nodes are established, objects with a specified type (e.g., MTDeviceType) are created within the server and variables with a specified data type are also created. Untyped nodes are converted to typed nodes that can be manipulated in the server, allowing data to be uploaded to these nodes. Any predefined nodes that remain untyped are replaced with their typed versions. This method of instantiation is extrapolated to devices to establish nodes for devices. For MTConnect devices, ideally the data should be streamed directly from the device to reduce the points of failure; however, in this study, due to time constraints, the data was streamed from a text file, which is generated by the MTConnect XML parser. Once the nodes are instantiated, a client can communicate with the server to access the information.

OPC-UA Client

Similar to the requirements for the server, the tool for the OPC-UA client development was selected based on its price and available documentation. Additional consideration included the programming language used, as clients come in many forms; some use GUIs while others are based in the programming environments of languages such as MATLab [10] or Python [11]. We used UaExpert by Unified Automation [12] because it facilitates the GUI usage, is free, and well documented. UaExpert enables the connection with the server and supports the testing plan because it has features, such as Data Access view, Data Logger view, and Event view. The comma-separated values (CSV) export feature in the Data Logger view enables data visualization through the GUI.

Graphical User Interface

To confirm proper transfer of data from the MTConnect device to the OPC-UA server/client, a GUI is developed to take in real-time data from the parser and client, to compare the two data sets, and display the plots. The GUI is developed using Dash by Plotly [13], which helps display data in real time and allows for expandable data sets and graphs that are particularly useful when dealing with multiple components or DataItems. The implementation uses a Dash module in Python, which requires data to be imported. The simplest method of doing so is through the text files. This method gives the user the option to select which graphs are visible based on the text files imported with a graph description given by the component and attribute names. The selected data set and its corresponding timestamp are imported and plotted in real-time as the text file continually grows with streaming data. Simplifications have been made to the GUI to reduce the memory usage on the system while promoting faster graph response time for recording since Dash GUIs are web-based interfaces displayed using porting on the local machine that can be accessed via a browser.

The GUI is developed for the visualization of highly dynamic data since UaExpert only displays a node's value, which can be changing constantly, sometimes faster than human eyes can detect, depending on the sampling rate. Attempting to compare rapidly changing position values of two data sets by eye does not uffice for validating a standard. The GUI is only needed for highly dynamic values since less dynamic values, such as events or conditions, are easy to assess visually; an Emergency Stop is either armed, triggered, or unavailable, and an oil temperature is either normal, unavailable, or in the warning or fault zone. These types of DataItems tend not to change repeatedly and therefore can be validated via observation, leveraging UaExpert's Data Access view since the value of a node is shown directly on its interface.

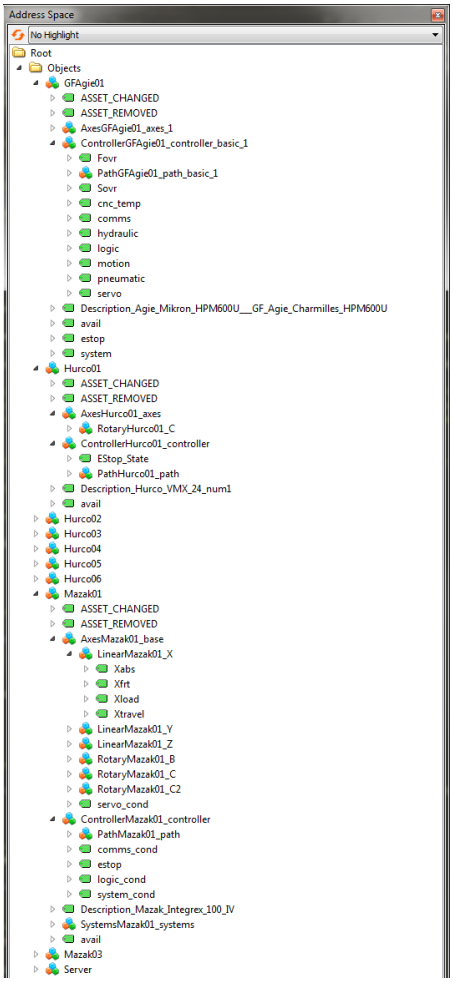


Figure 3. A Sample Partial NIST Test Bed Address Space Depicted in UaExpert

Fisher, Ryan; Shao, Guodong.
"TESTING OF THE MTCONNECT – OPC-UA COMPANION SPECIFICATION."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, Erie, PA, United States. June 10, 2019 - June 14, 2019.

TEST CASES

A variety of test cases must be defined to test the information mapping between the MTConnect agent and OPC server/client properly. Before defining test cases, we fist discuss the key features of OPC-UA and then briefly discuss the types of MTConnect data.

Key Features of UA

While OPC-UA has many features, Data Access (DA), Historical Access (HA), and Event and Alarm Notifiers (EA), are three typical ones that can be used to begin testing the MTConnect to OPC-UA gateway [3].

- The Data Access feature provides instant machine/process monitoring and control capabilities by allowing clients or applications to directly stream real-time data from the server.
- The Historical Access feature is essentially an extension of the Data Access feature, but it keeps record of previous values for specified nodes. Cases for using this feature include data analysis, machine failure prevention, modeling, and simulation.
- The Event and Alarm Notifiers component provides OPC-UA clients with the ability to detect a change in an event or condition. It triggers an alarm if such a change occurs for a pre-existing node. An example for effective use of this feature would be the initiation of an alarm for the MTConnect predefined EmergencyStop event that has three possible values, one of which is "Triggered," meaning an emergency stop has occurred. Setting an alarm for this type of event would notify applications and interfaces by relaying this important information. Alarms can be selected for specific events since machine operators may not desire to have alarms continually activating for each node defined as an event or condition.

These three core features on a specified test case are sufficient for determining if a proper mapping of information occurs.

Defining DataItem Types

There are three types of DataItems that can exist in an MTConnect model [14]:

(1) Samples: samples must be numeric values from a stream. Examples include Rotary Speed, Angle, and Position.
(2) Events: events can be a variety of data types and generally have a predefined controlled vocabulary for specific components; however, there are cases where only a character string representing data is returned by the device.
(3) Conditions: conditions are another type of DataItem that exist in the MTConnect model. Each component may have more than one condition active at a specific instance where the conditions are defined by the string type, but each condition can be in one of four states: Normal, Warning, Fault, or Unavailable.

From these DataItems as shown in Figure 4, test cases can be generated.

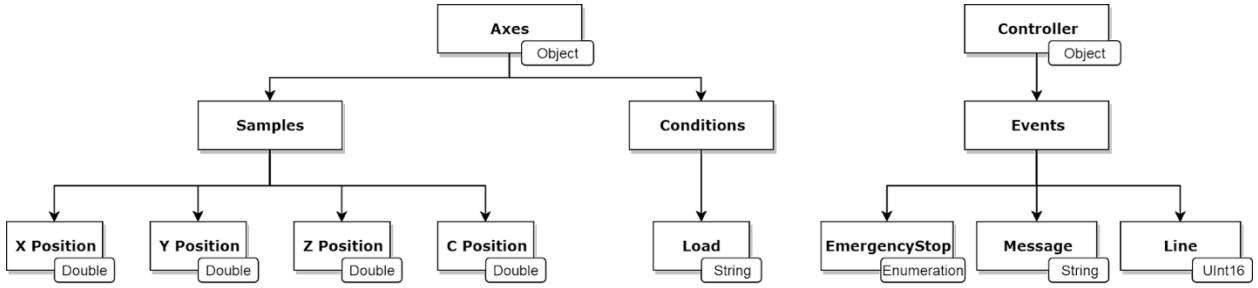


Figure 4. Types of DataItems

Defining Test Cases

The samples selected for test cases include the X, Y, Z, and C positions under the Axes component. A multitude of these DataItems were selected to ensure the server could handle multiple highly dynamic streams of data. For each of these four positions, the MTConnect schema or probe, given by the simulator, may display multiple DataItems such as an actual position, commanded position, and loaded position; the actual position is used for the mapping to a variable. The original MTConnect schema breaks down the Axes component into Linear Axis and Rotary Axis for better accuracy of data representation, however, a simplification is made by eliminating this extra component and placing the variables directly under the

Fisher, Ryan; Shao, Guodong.
"TESTING OF THE MTCONNECT – OPC-UA COMPANION SPECIFICATION."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, Erie, PA, United States. June 10, 2019 - June 14, 2019.

Axes component. If the Linear Axis and Rotary Axis were to be implemented in the testing, it would only be an additional object in the hierarchy of the node tree, viewed in the address space. This would make no difference in the testing of the mapping. Custom types (e.g., MTComponentTypes) could be created to map a device with a custom structure (e.g., Axes with only Rotary components), which allows for expansibility to other devices. The testing of custom types requires only a syntactic change in node generation executed by the probe.

The condition test case selected is the Load, which gives the load condition for the specified axis in the MTConnect schema (in our case the X Axis). This condition is selected since it pre-exists under the Axes component. This avoids the need to create another component such as "Coolant" or "Electric" to the address space. Note that only one condition was selected since conditions can only be a string type, and their output is one of four values.

The controller component is selected to host the event test cases since it provides a variety of data types under one object. The string and integer are standard data types, however, the enumeration requires an integer be passed to the server to access a string from the enumeration. To account for this, the parser must translate the XML string (e.g., Unavailable) to the corresponding integer in the enumeration definition. The enumeration definition for this variable could be defined in the specification or could be custom, depending on the device; prior knowledge of the enumeration list must be obtained. For all test cases, an assumption was made stating that each event DataItem would return its declared data type to define the rigidity of the data being returned by the machine. If an alternative data type is returned, it will not be processed, and the previous node value is kept.

Applying Test Cases

Each of the identified test cases and their corresponding DataItem is to be explicitly declared in the MTConnect files being passed into the XML probe and parser. The server accesses them via the .uanodes and stream data files generated respectively. A secure connection between the server and client enables the data to be properly transmitted. If the test case contains highly dynamic data, the GUI is used to access the parser and client for displaying the test cases and their data for I/O analysis. If the test case contains less dynamic data, then the initial XML file, Data Access view, and Event view (if applicable) are used to observe changes in data. Validation of the standard is performed via the I/O analysis.

TESTING RESULTS

Data Access

Using the simulator provided by NIST and MemexOEE, after probing the device structure for the test cases, the nodes were properly generated in the address space of the server for all components and DataItems specified on all the devices. This confirms the MTConnect XML probe operated correctly. The streaming of data from these devices into the server was completed successfully, as the client reported changing values with the Data Access feature. Using the Data Logger, the machine data for the X-Position was recorded and plotted directly from the machine simulator and the client via the GUI. The results shown in Figure 5 display the X-Position data over an approximate twenty-minute period. While there is no delay implemented into the parser, a one-hundred millisecond delay was established on the server.

Comparing the I/O of X-Position data, a conclusion can be made that the MTConnect data was properly transferred from the agent to the OPC-UA server and client. While the overall data mapping is executed correctly, there were small errors in parts of the stream that were foreseen. Due to the XML parser outputting the streamed data to a text file and the server extracting the data from the text file simultaneously, occasionally the server would be incapable of opening the file. To allow the server to continue operating, try-catch statements were implemented, which returned the previous value of the node as the current instantiation.

Issues found: The samples for the positions of the Axes component in the companion specification are currently defined as a string data type when being mapped to the server, even though MTConnect specifically defines their position data as numerical values only. This indicates a design issue in the specification. This definition discrepancy created an issue when attempting to use the Data Logging tool to save data for the GUI's real-time plotting since the Data Logging tool needs to receive a double data type, not a string. For practical purposes, having numerical values stored as doubles or floats is justifiable while a representation in strings is not.

Alternate solution: To temporarily correct this definition error (assuming it will be permanently fixed in an updated version), the string data type was simply altered to a double in the predefined nodes, and the streamed value was also changed to a double; this change allows the use of Data Logging to transmit data for the visual representation in the GUI. All test cases assessed passed as compliant with the Data Access feature after data type definitions were corrected.

Historical Access

Issues found: Another issue occurred when executing the Historical Access feature on the identified test cases. Using UaExpert and its History view, the historical data was unable to be retrieved from the nodes. Attempts have been made to the nodes within the server to fix this error: the access levels were altered to allow history reading, history writing, and a current reading. The "Historizing" attribute was also enabled. The UaExpert still reported the node history to be empty. With the generation of this error, it is expected that nodes were not storing

Fisher, Ryan; Shao, Guodong.
"TESTING OF THE MTCONNECT – OPC-UA COMPANION SPECIFICATION."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, Erie, PA, United States. June 10, 2019 - June 14, 2019.

their historic values; the process of enabling the node storage functionality could not be determined.

Alternate solution: Ideally this should be engaged automatically in the generation of the node structure to eliminate the need for altering individual node attributes within the server. After using a C# .NET client to confirm this error, the Historical Access feature worked to a small extent by recording and storing values only after the client was started (similar to the Data Logger). This limited functionality, however, is attributed to the client's capabilities and therefore is not sufficient for justifying full usage of the Historical Access feature. Once the nodes are capable of storing their historic values, the History Access functionality is expected to be operable. To modify the nodes for this capability, the standard developers must investigate it further and address the issue accordingly.

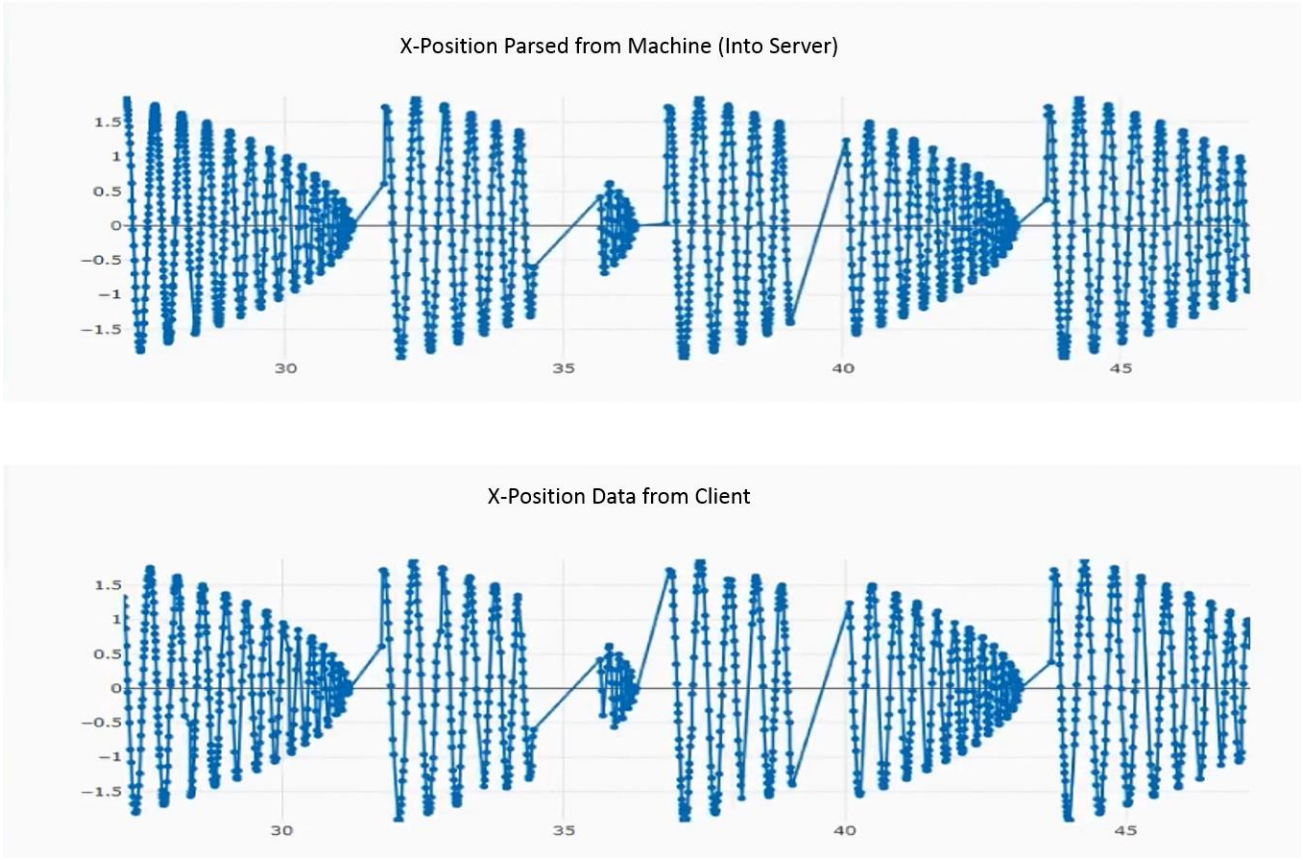


Figure 5. X-Position Data from the Machine and the OPC-UA Client

Event and Alarm Notifiers

Issues found: Another key issue that occurred during the testing was the use of the Event and Alarm Notifiers feature. From the defined test cases, the condition and event nodes were unable to respond properly to the feature. Although we tried different data types and used an alternative client, the feature continued to fail. We also modified the primary object's (device's) "EventNotifier" attribute by establishing a "SubscribeToEvents" setting. This revision did not allow the Event and Alarm Notifiers feature to recognize any changes in the event test cases. The same modification was also applied to the sub-objects (sub-components) of the primary object, and no changes occurred. The absence of the feature may be due to improper configuration of the server, an improper mapping of the companion specification, and/or a lack of description in the specification; more investigation should be performed.

CONCLSUIONS

In this paper, we discussed an initial implementation and testing of the MTConnect - OPC-UA Companion Specification

Fisher, Ryan; Shao, Guodong.
"TESTING OF THE MTCONNECT – OPC-UA COMPANION SPECIFICATION."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, Erie, PA, United States. June 10, 2019 - June 14, 2019.

Version 1.02. Our initial testing results show some issues in the specification that need to be addressed in future versions of the standard. These improvements will help the companion specification to better integrate semantics of MTConnect with the syntactic representation of OPC-UA. By conducting an expanded investigation of this and future versions of the standard, the specification could eventually be established as a viable use in the smart manufacturing setting.

The following recommendations regarding data types, Event and Alarm Notifiers, and enumerations, can be made to the SDOs based on the conducted testing.

- Data type definitions should be improved to properly map data for practical uses, i.e., the positions being mapped as doubles or floats compared to strings.
- The enablement of Event and Alarm Notifiers should be explicitly declared in newer versions of the specification, as currently there is no method stating how objects are enabled as events. Whether the enablement of nodes for this feature occurs within the server or within the model generation, the companion specification should state one or both options. A simple solution to be proposed is enabling Event and Alarm Notifiers upon generating nodes, allowing the OPC-UA server to establish a clear distinction between imported DataItems. If enablement of notifiers is to be completed within the server, it will require a tedious task for users with complex device structures.
- Regarding enumerations, for data to properly be retrieved by the server/client, an integer representing position of the data type in the enumeration had to be imported into the server. While this justifies the functionality of the companion specification, it must be noted that the data taken from the stream (string retrieved such as “ARMED”) had to be converted to an integer representing the proper position in the enumeration. An example of this is the streamed value of “ARMED” being translated to the integer “1”, which represents the second position in the enumeration list declared in the MTConnectModel.xml. The integer is then used by the server to locate the proper string value. This requirement means that a converter must be made for all enumerations to allow the server to properly access data using the current method. While this may or may not be a correction that needs to be accounted for in future versions, it is an aspect of the companion specification that should be brought to the SDOs’ attention.

Future work for better validating the standard includes increasing the number of test cases, performing a larger variety of test cases, if possible, in which further examination of data types is needed, as other errors are bound to exist, similar to the position data type error. To correct for the minor errors that were occurring in the data streaming process, the parser should be implemented using C# so that it can execute directly within the server. By having the server access the agent directly, compared to indirectly through text files, the chances of data-access errors occurring will be reduced or eliminated since files will not be opened and closed simultaneously. Finally, we should work closely with the standard developers to figure out the issues with the testing of the Event and Alarm Notifiers feature.

NOMENCLATURE

CSV Comma-Separated Values
DA Data Access
DX Data Exchange
EA Event and Alarm notifiers
GUI Graphical User Interface
HA Historical Access
NIST National Institute of Standard and Technology
OPC Object linking and embedding for Process Control
SDO Standard Developing Organizations
SM Smart Manufacturing
UA Unified Architecture
URL Uniform Resource Locator
V&V Verification and Validation
XML Extensible Markup Language

ACKNOWLEDGMENTS

The authors would like to thank the Summer Undergraduate Research Fellowship (SURF) program for supporting the project and thank Moneer Helu, Will Sobel, and Randy Armstrong for their valuable discussion and support.

DISCLAIMERS

No approval or endorsement of any commercial product by the National Institute of Standards and Technology (NIST) is intended or implied. Certain commercial software systems are identified in this paper to facilitate understanding. Such identification does not imply that these software systems are necessarily the best available for the purpose.

REFERENCES

[1] MTConnect (2018) A free, open standard for the factory. http://www.mtconnect.org
[2] OPCUA (2018) Unified Architecture. https://opcfoundation.org/about/opc-technologies/opc-ua/.
[3] MTConnect (2013) MTConnect - OPC UA Companion Specification Release Candidate Version 1.02.
[4] Hirvonen, Markus. “Streamlining Manufacturing Data Integration.” Tampere University of Technology, Tampere, Finland. 2017. https://core.ac.uk/download/pdf/144141231.pdf
[5] Memex (2018) Driving efficiency and productivity from the shop floor to the top floor. http://www.memexoee.com/.
[6] Urllib (2018) Open arbitrary resources by URL. https://docs.python.org/2/library/urllib.html.

Fisher, Ryan; Shao, Guodong.
”TESTING OF THE MTCONNECT – OPC-UA COMPANION SPECIFICATION.”
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, Erie, PA, United States. June 10, 2019 - June 14, 2019.

[7] ElementTree (2018) The ElementTree XML API. https://docs.python.org/2/library/xml.etree.elementtree.html.
[8] UA-ModelCompiler (2018) Model compiler converts XML files into C# and ANSI C. https://github.com/OPCFoundation/UA-ModelCompiler.
[9] OPCUA .NET (2016) Build OPC UA .NET applications using .NET Standard Library. http://opcfoundation.github.io/UA-.NETStandard/.
[10] MathWorks (2018) MATLAB. https://www.mathworks.com/products/matlab.html.
[11] Python (2018) A programming language that lets you work quickly and integrate systems more effectively. https://www.python.org.
[12] UaExpert (2018) A full-featured OPC UA Client. https://www.unified-automation.com/products/development-tools/uaexpert.html.
[13] Dash Plotly (2018) Build beautiful web-based interfaces in Python. https://plot.ly/products/dash/.
[14] MTConnect (2014) MTConnect Specification and Materials. http://www.mtconnect.org/docs/streams/.

Fisher, Ryan; Shao, Guodong.
"TESTING OF THE MTCONNECT – OPC-UA COMPANION SPECIFICATION."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, Erie, PA, United States. June 10, 2019 - June 14, 2019.

Long-Slotted Plate Connections for Enhancing the Robustness of Steel Gravity Systems against Column Loss: Preliminary Results

J. M. Weigand[1*], T. Thonstad[1], and A. A. Seamone[2]

[1]Research Structural Engineer, Engineering Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8611
[2]Undergraduate Research Fellow, Civil, Environmental and Architectural Engineering, University of Colorado Boulder, 1111 Engineering Drive, Boulder, CO 80309-0428
[*]Corresponding author. Tel: (301) 975-3302; E-mail: jonathan.weigand@nist.gov

ABSTRACT

Steel gravity frames are commonly used in United States building construction practice, but they are potentially vulnerable to disproportionate collapse under column loss, as has been shown by recent experimental and analytical studies. To overcome these vulnerabilities, a new type of connection has been developed. These "enhanced" gravity connections, which could be implemented in new or existing structures, incorporate long-slotted steel plates that are welded to the column and bolted to the top and bottom flanges of the beam. An experimental program designed to evaluate the influence of the key geometric factors on the coupled flexural-axial performance of these enhanced connections is underway. This paper presents selected results from the first block of the full experimental design: the behavior, failure mode, and key measured quantities from a set of eight tests conducted on replicate specimens with nominally identical geometry. Long-slotted plates were axially tested in a single-lapped bolted configuration under monotonic tensile loads, to characterize the behavior and failure mode of the components. The preliminary results show that the connection bolts maintain pretension while slipping within the slot, and do not lose appreciable pretension until bearing occurs. The results were highly repeatable; the coefficient of variation in the peak tensile load, and displacement at the peak tensile load was 0.025 and 0.032, respectively.

INTRODUCTION

Recent research has identified that steel gravity framing systems with conventional shear-tab connections are potentially vulnerable to disproportionate collapse under column loss scenarios. Four column removal tests performed on a half-scale steel gravity framing system with composite slab on steel deck showed that the floor system could only carry between 44 % and 62 % of the applicable gravity load combination [Johnson et al. 2015]. To help address this potential vulnerability, "enhanced" steel gravity connections that incorporate U-shaped top and seat plates, with long-slotted holes bolted to the beam flanges, have been developed. These enhanced connections could be implemented in new construction, or as a retrofit strategy for existing structures. The enhanced connections have been shown, using computational modeling, to provide more than double the resistance of the conventional shear-tab

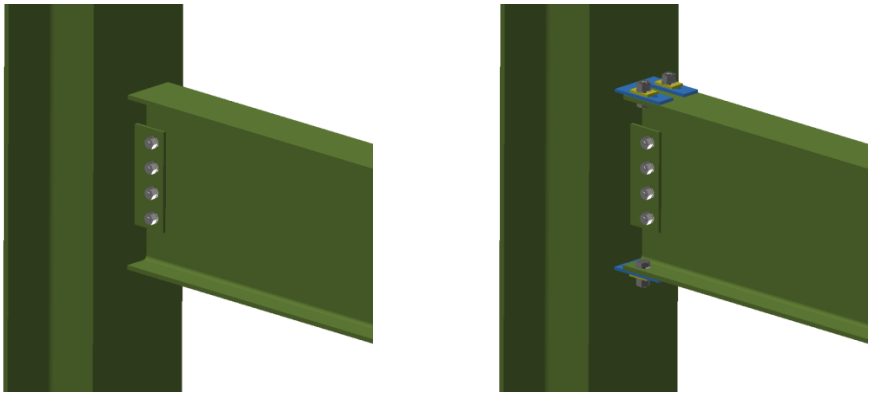
Weigand, Jonathan; Thonstad, Travis; Seamone, Andrew.
"Long-Slotted Plate Connections for Enhancing the Robustness of Steel Gravity Systems against Column Loss: Preliminary Results."
Paper presented at Structures Congress 2019, Orlando, FL, United States. April 24, 2019 - April 27, 2019.

connections [Weigand 2014; Weigand and Main 2016] when subjected to combined rotational and axial loads consistent with column removal. When implemented in system-level analyses of a two-bay by two-bay composite floor system, the enhanced connections increased the vertical load-carrying capacity of the system under center column loss by 90 % under static loading [Weigand and Main 2016]. These analyses also indicated that the system with the enhanced connections has the potential to resist the applicable gravity load combination under instantaneous dynamic center column loss. To begin developing a widely applicable design procedure, experimental validation and optimization of the U-shaped top and seat plates are needed.

This paper presents selected, preliminary results from an experimental program designed to evaluate the influence of geometric factors on the coupled flexural-axial performance of these enhanced connections. The intent of this research is to determine the optimal configuration of the slotted U-shaped plates for the range of anticipated connection configurations used in practice. Single-lapped long-slotted plate specimens were axially tested in a single-lapped bolted configuration under monotonic tensile loading, to-characterize their behavior and failure modes.

ENHANCED CONNECTIONS

An enhanced steel gravity connection is shown in Fig. 1. The connection includes a conventional shear-tab connection and top and seat plates, which are welded to the column and bolted to the beam flanges. Long-slotted holes in the top and seat plates permit large axial slip displacements of the flange bolts to occur prior to bearing at the ends of the slots. Rectangular plate washers distribute the bearing stresses induced by pre-tension in the flange bolts. The net section of the plate is designed relative to the shear area of the bolts to ensure significant plastic deformations are achieved in the plate and/or bolt prior to rupture, and can also be capacity-designed so that plate net tensile rupture over bolt shear rupture is the controlling failure mode. Standard holes are used in the beam flanges.



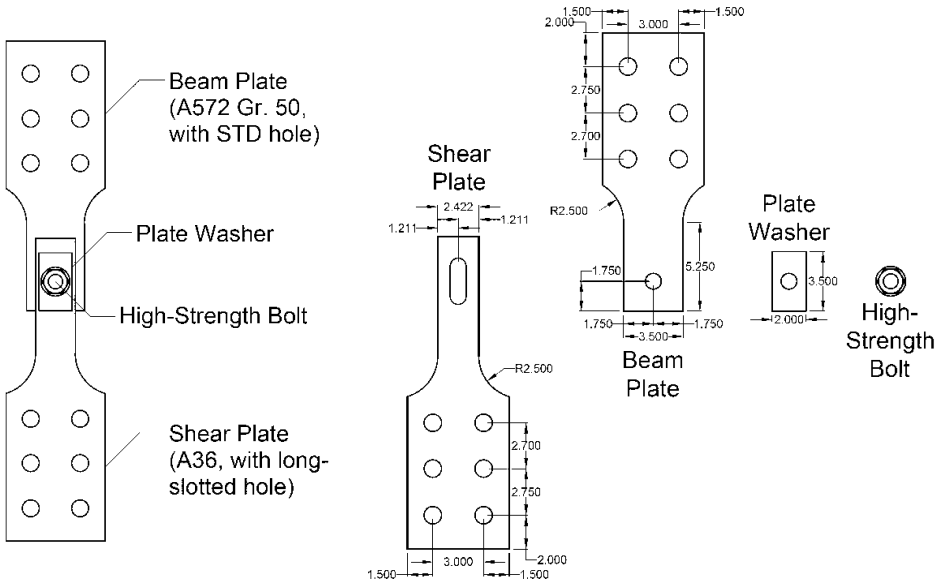
(a) (b)

Fig. 1. (a) Conventional single-plate shear connection, and (b) enhanced single-plate shear connection incorporating U-shaped slotted top and seat plates welded to column

Weigand, Jonathan; Thonstad, Travis; Seamone, Andrew.
"Long-Slotted Plate Connections for Enhancing the Robustness of Steel Gravity Systems against Column Loss: Preliminary Results."
Paper presented at Structures Congress 2019, Orlando, FL, United States. April 24, 2019 - April 27, 2019.

COMPONENT TESTS

Component tests were performed to characterize the influence of the geometry of the top and seat plates on the coupled flexural-axial performance of the enhanced connection concept. The design of the component test specimens reduced the U-shaped top and seat plates down to their simplest possible component parts: representative widths of the top or seat plate with a long-slotted hole and beam-flange plates with a standard hole, as shown in Fig. 2. The two plates were connected by a single high-strength bolt. A 7.9 mm (5/16 in) thick plate washer covered the slot to distribute the bearing load induced by the bolt pretension over a larger area. To be as representative as possible of typical steel construction practices, the shear plates were fabricated from ASTM A36 steel [ASTM, 2014], while the beam-flange plates were fabricated from ASTM A572 Grade 50 steel [ASTM, 2018a]. By reducing the complexity of the test specimen, the influence of the geometric parameters of the components and the interactions between these parameters could be efficiently studied.



(a) (b)

Fig. 2: (a) typical steel component specimen, and (b) exploded view with annotated dimensions. All dimensions in inches, 1 in = 25.4 mm.

The design of experiments was based on a central composite rotatable design. The selected configurations were used to examine the influence of five aspects of the connection geometry, including the load ratio (i.e., the ratio of the bolt single-shear strength to the tensile rupture strength of the slotted plate), slot length ratio (i.e., slot length relative to the long-slot length (LSLT) in the Steel Construction Manual [AISC, 2017]), aspect ratio of the plate legs adjacent to the slot (width of leg to thickness of shear plate), bolt diameter, and bolt pretension ratio (i.e., ratio of the bolt pretension to $1.13T_{min}$, where $1.13T_{min}$ is the average installed bolt pretension, and T_{min} is the minimum bolt pretension specified in the Steel Construction Manual [AISC, 2017]). This paper presents selected results from the first block of the full

experimental design: the behavior, failure mode, and key measured quantities from a set of eight tests conducted on replicate specimens with nominally identical geometries (see Table 1).

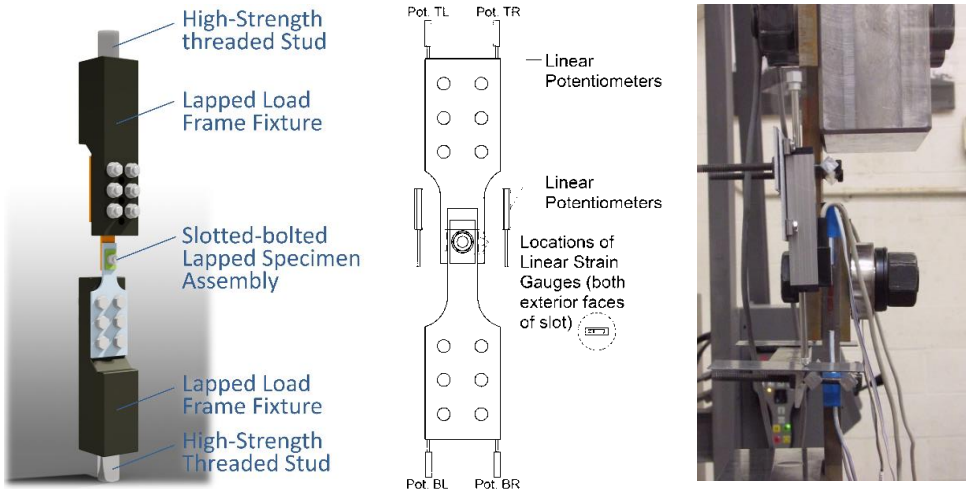
Table 1: Specimen geometric parameters

Specimen ID	Bolt Diameter, mm (in)	Load Ratio	Slot Length Ratio	Bolt Pretension Ratio	Bolt Pretension kN (kip)	Aspect Ratio
SP001 SP002 SP003 SP004 SP005 SP006 SP007 SP008	22.2 (0.875)	1.10	1.25	1.00	196 (44.1)	1.50

The test protocol involved raising the upper fixture that holds the end of the beam plate, so that the lapped connection was placed under tensile loading. The tests were conducted in a self-reacting 1000 kN (220 kip) capacity servohydraulic load frame. The total expanded uncertainties in the measured actuator displacement and load were ±0.38 mm (0.015 in and ± 2.6 kN (0.6 kips), respectively. More information about how these uncertainties were evaluated is given in Appendix A. An impact wrench was used to tighten the six-bolt connections of the steel plates to the test fixtures to minimize any unanticipated displacements between the specimen and test fixtures during the tests. The specimen bolt was tightened by hand to the specified target pretension, measured using a load cell washer. The total expanded uncertainty in the measured bolt pretension was ±3.6 kN (0.8 kips). The tests were pseudo-static and were conducted in displacement control at a constant displacement rate of 2 mm/min (0.075 in/min).

The instrumentation for the component tests is shown schematically in Fig. 3b. Six potentiometers and four strain gauges were mounted on the specimens to measure relative displacements throughout the test rig and deformations within the specimen. These measurements were made to verify that the recorder displacements corresponded to deformations within the slotted connection region of the specimens. Four potentiometers were used to measure the slip between the test fixtures and the specimen, and two potentiometers were placed on either side of the bolted central connection. The total expanded uncertainty of these measurements was ±0.6 mm (0.02 in) (See Appendix A). The four strain gauges were placed on the side of the A36 slotted plate (two on each leg) to measure the deformation of the legs surrounding the slot. A load cell washer on the back side of the beam-flange plate allowed for the tension in the bolt to be measured directly.

Weigand, Jonathan; Thonstad, Travis; Seamone, Andrew.
"Long-Slotted Plate Connections for Enhancing the Robustness of Steel Gravity Systems against Column Loss: Preliminary Results."
Paper presented at Structures Congress 2019, Orlando, FL, United States. April 24, 2019 - April 27, 2019.



(a) (b) (c)

Fig. 3: (a) test configuration, and (b) instrumentation (c) instrumented specimen

OBSERVED RESULTS

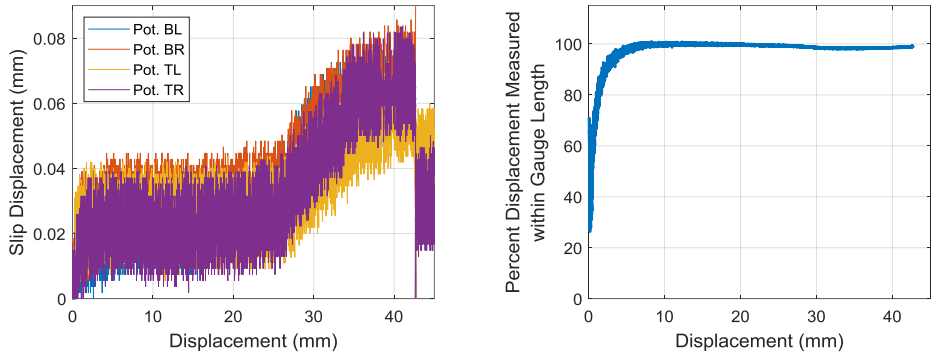
Visible polishing of the surfaces of the lapped steel plates occurred during each test. These polished surfaces suggest that the coefficient of friction between the surfaces decreases throughout the test. Due to the inherent eccentricity in the lapped connection, significant bending of the two lapped plates and rotation of the connection bolt occurred once the bolt made contact and with the end of the slot, but not before. Each of the eight replicate specimens experienced bolt, rather than plate, rupture. This was anticipated, since a load ratio of greater than unity indicates that the plate strength was designed to be stronger than the bolt shear strength. The sheared bolts had a shiny crescent across the diameter of their cross-sections, as shown in Fig. 4. These shiny and dull lusters of the bolt rupture surface are attributable to slow, and fast fracture rates, respectively. Although standard procedures were used in applying the strain gauges, they did not remain attached throughout the entire test due to localized rotations occurring in the legs of the slotted shear plate. The strain gauges could typically only reach strains of 3 % to 4 % before becoming detached.

Fig. 4: shear rupture in bolt (SP005)

MEASURED RESULTS

Slip Displacements

The potentiometer measurements from the tests were used to verify that displacements in the test setup were concentrated within the lapped connection region of the specimen and minimized elsewhere. Fig. 5a shows the measured slip vs. actuator displacement for the four slip potentiometers for a typical test (see Fig. 3(b) for naming convention). Each potentiometer behaved similarly throughout the tests and did not reach a slip value of greater than 0.10 mm (0.0040 in) for any test. Given that the measured slips correspond to less than 0.3 % of the total deformation of the specimen at peak load, slip displacements were deemed negligible, and are subsequently ignored.



(a) (b)

Fig. 5: (a) slip displacements, and (b) percent displacement measured within gauge length

The total measured displacement, the sum of the slip deformation in the test setup and the deformation of the test specimen, was compared to the displacement measured by the hydraulic actuator's Linear Variable Differential Transformer (LVDT). Fig. 5(b) shows the ratio of the measured displacement within the lapped connection region to the actuator displacement as a percentage. This curve corresponds to the test with largest difference between the instrumentation and the actuator data. After the actuator displaced roughly 5 mm (0.2 in), the measured displacement within the lapped connection region was consistently within 1 % to 3 % of the measured actuator displacement. This is consistent with the negligible slip displacements measured between the test fixtures and the specimen. The good agreement between the actuator displacement and the displacement measurements gives confidence that the test configuration behaved as intended.

Force Displacement Response

Fig. 6 shows the typical force-displacement behavior of one of the tested specimens. Vertical lines designate four, distinct phases of the typical force-displacement response. Phase I represents the elastic portion of the connection response. This phase is typified by elastic deformations, with an initial stiffness of k_i, within the plates before

slipping of the bolt starts to occur along the slot. Phase II represents the bolt-slip portion of the response, during which the bolt travels along the length of the slot. The slope of the slip portion of the response is denoted as k_y. Phase III begins when the bolt has traveled the length of the slot and the bolt shaft comes into full bearing contact with both the end of the slot and the standard hole in the beam plate. This region is designated as the bearing stage of the connection response. The initial slope at the transition to Phase III is denoted as k_b. Phase IV begins after the specimen reaches its ultimate load, and failure of the connection begins to occur. The failure mode determines how abruptly the connection fails. Specimens that fail due to bolt shear, as was the case for the eight tested nominally identical specimens, lose strength abruptly, as demonstrated by Fig. 6.

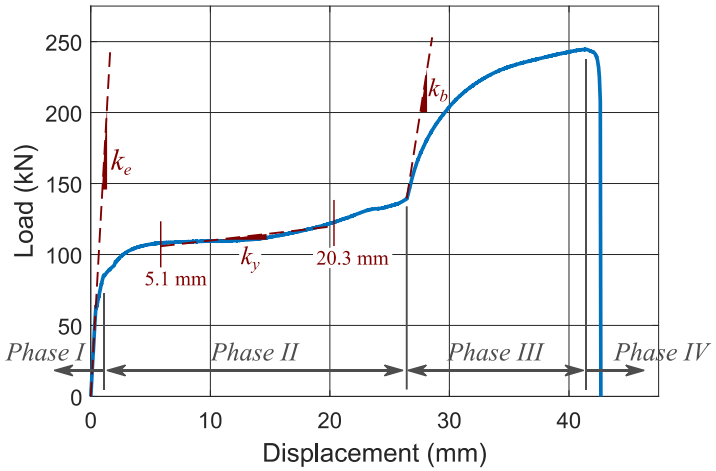


Fig. 6: Phases of the typical specimen load-displacement response

Bolt tension

A load cell washer was used to measure the pretension in the connection bolt throughout each test. The load in the load washer during testing of SP005 is shown in Fig. 7 along with the specimen's load-deformation response for context. The bolt maintained its pretension through the entire slip phase of the test and did not lose appreciable pretension (i.e. more than 15 %) until bearing occurred. Once the bolt came into bearing with the steel surface, the measured tension in the load cell washer decreased significantly; although some pretension remained in the bolt even at rupture.

Weigand, Jonathan; Thonstad, Travis; Seamone, Andrew.
"Long-Slotted Plate Connections for Enhancing the Robustness of Steel Gravity Systems against Column Loss: Preliminary Results."
Paper presented at Structures Congress 2019, Orlando, FL, United States. April 24, 2019 - April 27, 2019.

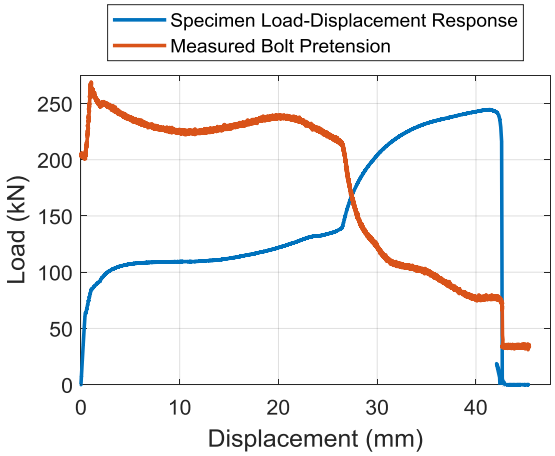


Fig. 7: Bolt tension versus displacement for SP005, which failed due to bolt shear rupture

VARIATION IN MEASURED RESPONSE

Table 2 shows key measured quantities of the force-displacement responses of the eight nominally identical test runs. The maximum force and its corresponding displacement are reported, along with the three stiffness values that characterize the shape of the load-displacement curve. The parameter, k_e, represents the initial elastic stiffness of the connection prior to slip (Phase I in Fig. 6); k_y, the average tangent slope of the slip region between 5.1 mm (0.20 in) and 20.3 mm (0.80 in); and k_b, the initial slope at initiation of bearing (beginning of Phase III).

Table 2: Center Point Test Data

			Stiffness Value		
Specimen ID	**Peak Load,**	**Displacement at Peak Load,**	k_e	k_y	k_b
	kN (kip)	mm (in)	kN/mm (kip/in)	kN/mm (kip/in)	kN/mm (kip/in)
SP001	265 (59.5)	38.1 (1.50)	147 650 (843.1)	1874 (10.7)	29 260 (167.1)
SP002	260 (58.5)	38.6 (1.52)	141 470 (807.8)	2644 (15.1)	33 480 (191.2)
SP003	256 (57.6)	40.6 (1.60)	135 600 (774.3)	1979 (11.3)	41 770 (238.5)
SP004	250 (56.1)	41.7 (1.64)	137 190 (783.4)	1944 (11.1)	36 430 (208.0)
SP005	245 (55.0)	41.4 (1.63)	134 460 (767.8)	1261 (7.2)	30 400 (173.6)
SP006	250 (56.1)	39.9 (1.57)	143 690 (820.5)	1594 (9.1)	23 940 (136.7)
SP007	250 (56.3)	41.7 (1.64)	131 030 (748.2)	1173 (6.7)	27 650 (157.9)
SP008	247 (55.6)	39.9 (1.57)	149 490 (853.6)	1716 (9.8)	44 590 (254.6)
Mean	253 (56.8)	40.1 (1.58)	140 070 (799.8)	1769 (10.1)	33 430 (190.9)
COV	0.025	0.032	0.044	0.244	0.199

At failure, the specimens with nominally identical geometries had a mean peak resistance of 253 kN (56.8 kip), with a coefficient of variation of 0.025. The small coefficient of variation in the ultimate load indicates that the results were highly

repeatable. The mean initial stiffness of the connection was 140 070 kN/mm (799.8 kip/in). This represents the elastic stiffness of the steel connection before slipping occurs. The mean stiffness of the slip response was 1769 kN/mm (10.1 kip/in). The mean initial stiffness of the bearing stage of the experiment was 33 430 kN/mm (190.9 kip/in). This represents the rate at which the connection reaches its maximum load before failure.

SUMMARY AND PRELIMINARY CONCLUSIONS

This paper presented selected results from an experimental program designed to characterize the influence of key geometric factors on the coupled flexural-axial performance of an enhanced connection concept for steel gravity framing systems. Long-slotted plates were axially tested in a single-lapped bolted configuration under monotonic loads to characterize the behavior and failure modes of the components.

From the component testing program, several preliminary conclusions can be drawn:

- Displacements in the test setup were concentrated within the lapped connection region of the specimen, as intended. Slip between the test fixtures and the specimen was negligible, accounting for less than 0.3 % of the measured displacement at the maximum force in each of the tests.

- Bolts maintained more than 85 % of their pretension through the entire slip phase of the tests. The bolts did not lose appreciable pretension until bearing occurred. Once the bolt shank came into bearing with the inner surfaces of the holes in the steel plates, the tension measured by the load cell washer decreased abruptly prior to specimen failure

- The tests were repeatable. Key measured quantities of the force-displacement responses of the eight nominally identical specimens were similar. The coefficient of variation (COV) for the maximum force was 0.025. The largest variations were in the post-slip and bearing stiffnesses: the COV for these measurements was 0.244 and 0.199, respectively.

DISCLAIMER

ACKNOWLEDGEMENTS

A portion of this work was performed by Andrew Seamone during his tenure as a NIST Summer Undergraduate Research Fellow (i.e., SURF student). The authors would like to thank the NIST SURF Program for supporting Mr. Seamone's contribution to this work.

REFERENCES

AISC. (2011). Steel construction manual, 14 Ed., Chicago.

ASTM (2014). "Standard Specification for Carbon Structural Steel." Standard A36/A36M-14, ASTM International, West Conshohocken, PA.

ASTM (2018a), "Standard Specification for High-Strength Low-Alloy Columbium-Vanadium Structural Steel." Standard A572/A572M-18, ASTM International, West Conshohocken, PA.

Johnson, E.S. Meissner, and Fahnestock, L.A. (2015). "Experimental Behavior of a Half-Scale Steel Concrete Composite Floor System Subjected to Column Removal Scenarios." *Journal of Structural Engineering*, 04015133.

Taylor, B. N., and Kuyatt, C. E. (1994). "Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results." NIST Technical Note TN-1297, Gaithersburg, MD.

Weigand, J. M. and Main, J. W. (2016) "Enhanced Connections for Improved Robustness of Steel Gravity Frames." *Proc., Eighth International Workshop on Connections in Steel Structures (Connections VIII)*, Boston, Massachusetts, May 2016.

Weigand, J.M. (2014). "The Integrity of Steel Gravity Framing System Connections Subjected to Column Removal Loading." Ph.D. Dissertation in Civil Engineering, University of Washington, Seattle, WA.

Weigand, Jonathan; Thonstad, Travis; Seamone, Andrew.
"Long-Slotted Plate Connections for Enhancing the Robustness of Steel Gravity Systems against Column Loss: Preliminary Results."
Paper presented at Structures Congress 2019, Orlando, FL, United States. April 24, 2019 - April 27, 2019.

APPENDIX A: UNCERTAINTY IN MEASUREMENTS

The measurements presented in this document include length (deformations and position) and force (applied load). For each measurement and instrument variety, Type A and/or Type B uncertainties, combined standard uncertainties, and total expanded uncertainties were estimated. As defined in Taylor and Kuyatt (1994), Type A uncertainty was evaluated using statistical methods; Type B uncertainty was estimated by other means such as the information available in manufacturer's specifications, from past-experience, or engineering judgement. The combined standard uncertainty was estimated by combining the individual uncertainties using "root-sum-of-squares" (Taylor and Kuyatt, 1994). The expended uncertainty was then computed by multiplying the combined uncertainty by a coverage factor of 2 corresponding to an approximately 95 % confidence interval.

Table A-1 summarizes the components of the measurement uncertainty. All uncertainties are assumed to be symmetric (+/-) and to have a Gaussian distribution.

Table A-1: Measurement Uncertainty

Measurement/Component	Type	Component Standard Uncertainty	Combined Standard Uncertainty	Total Expanded Uncertainty (k=2)
Actuator position				
Uncertainty in secondary standard	B	0.2 mm (0.006 in)	0.2 mm (0.007 in)	0.4 mm (0.015 in)
Uncertainty in calibration procedure (N=32)	A	0.2 mm (0.004 in)		
Actuator load				
Uncertainty in secondary standard	B	1.3 kN (0.3 kips)	1.3 kN (0.3 kips)	2.6 kN (0.6 kips)
Uncertainty in calibration procedure (N=32)	A	0.4 kN (0.1 kips)		
Load cell washer				
Uncertainty in secondary standard	B	1.3 kN (0.3 kips)	1.8 kN (0.4 kips)	3.6 kN (0.8 kips)
Uncertainty in calibration procedure (N=7)	A	1.3 kN (0.3 kips)		
Displacement Transducers				
Uncertainty in secondary standard (N=8)	A/B	2 µm (0.00006 in)	0.3 mm (0.01 in)	0.6 mm (0.02 in)
Uncertainty in calibration procedure (N=20)	A	0.3 mm (0.01 in)		

Weigand, Jonathan; Thonstad, Travis; Seamone, Andrew.
"Long-Slotted Plate Connections for Enhancing the Robustness of Steel Gravity Systems against Column Loss: Preliminary Results."
Paper presented at Structures Congress 2019, Orlando, FL, United States. April 24, 2019 - April 27, 2019.

Wireless Interference Estimation Using Machine Learning in a Robotic Force-Seeking Scenario

Richard Candell*, Karl Montgomery*, Mohamed Kashef†, Yongkang Liu†
*Intelligent Systems Division †Advanced Network Technologies Division
National Institute of Standards and Technology (NIST)
Gaithersburg, Maryland, United States
Email: {richard.candell, karl.montgomery, mohamed.kashef, yongkang.liu}@nist.gov

Sebti Foufou
Computer Science Department
University of Burgundy
Dijon, France
sfoufou@u-bourgogne.fr

***Abstract*—Cyber-physical systems are systems governed by the laws of physics that are tightly controlled by computer-based algorithms and network-based sensing and actuation. Wireless communication technology is envisioned to play a primary role in conducting the information flows within such systems. A practical industrial wireless use case involving a robot manipulator control system, an integrated wireless force-torque sensor, and a remote vision-based observer is constructed and the performance of the cyber-physical system is examined. By using readings from the remote observer, an estimation system is developed using machine learning regression techniques. We demonstrate the practicality of combining statistical analysis with machine learning to indirectly estimate signal-to-interference of the wireless communication link using measurements from the remote observer. Results from the statistical analysis and the performance of the machine learning system are presented.**

***Index Terms*—industrial wireless, 802.11, factory communications, cyber-physical systems, wireless networking, robotics**

I. Introduction

The advances in wireless devices for cyber-physical systems (CPS) have led to rapid adoption of the industrial wireless system (IWS) in factories. The use cases for IWSs include process monitoring and control, discrete manufacturing, safety systems, and flexible factory work cells [1]. Implementation of wireless systems for industry has many advantages due to the lower cost, ease of scale, and flexibility due to the absence of cabling. However, these advantages come with challenges [2]–[4]. Such challenges include unpredictable latency, error uncertainty, and increased information loss when operating in the presence of significant interference and limited spectral resources [5]. When network operations are impaired by interference, fading, and propagation loss, the physical system performance may also be impaired. These impairments contribute to a change in the quality of information flow between wireless nodes and require careful co-design of the network and controller [6], [7].

Interference presents a significant challenge to IWSs and the underlying physical systems that rely on them. While wireless systems can be designed to support many users and devices and cognitive radio can be useful in scheduling transmissions and avoiding problem locations within the spectrum, sometimes it is not possible to avoid interference entirely or responses to interference is too slow. Sources of interference may be narrowband or wideband in nature. They include multi-path reflections, competing wireless systems, non-communication devices such as microwave ovens and industrial machines, and intentional jamming [8].

Methods to estimate, avoid, or mitigate interference are required for the deployment of reliable and deterministic IWSs. Existing methods rely on traditional signal processing and novel cognitive radio techniques. In [9], a method of desensitizing a 5G cellular network using interference cancellation of transmissions from neighboring cells is presented. Interference cancellation equipment is highly complex and costly, and the impact to latency in a CPS must be well understood. In [10], a method using a dedicated link quality estimation (LQE) node using received signal strength and information (RSSI) obtained from received data packets to identify interference and multi-path is presented as a viable approach to LQE in IEEE 802.15.4 networks without introducing additional traffic. In [11], a taxonomy of channel link quality techniques is presented providing a valuable survey on LQEs and asserting importance of link quality estimation in IWSs. In [12], failure analysis and wireless network troubleshooting are performed whenever the CPS is not functioning properly. Interference analysis is one major part of the troubleshooting procedure which is performed through traffic patterns and wireless spectrum analysis. Also, in [13], the use of spectrum analysis for interference detection and estimation is proposed for IWSs.

Candell, Richard; Montgomery, Karl; Hany, Mohamed; Liu, Yongkang; Foufou, Sebti.
"Wireless Interference Estimation Using Machine Learning in a Robotic Force-Seeking Scenario."
Paper presented at 28th International Symposium on Industrial Electronics (ISIE), Vancouver, Canada. June 12, 2019 - June 14, 2019.

LQE is one important but insufficient aspect of assessing the impact of link quality on a CPS. We assert that by jointly observing the performance of the physical and wireless components of a CPS, one obtains the complete perspective of the quality of the wireless link and its impact on physical performance. Since interference is such an important topic in the wireless CPS, we are motivated to contribute a method that simultaneously (1) makes observations of the physical system using ground truth measurements, and (2) infers the quality of the wireless communication system in terms of signal-to-interference ratio (SIR) using a relevant use case found in industry.

In this paper, we present a method using random forest regression to estimate the SIR ratio of the communication channel within a robotic arm force-seeking scenario in which the force signal is transmitted over a wireless local area network (WLAN) [14]. Position data from a vision-based tracking system, a distant observer, is used to train a channel quality estimator to infer the SIR of the wireless channel. The experiment is designed so that the small perturbations in the wireless channel resulting from interference will present position uncertainty in the physical system.

Our paper is organized as follows: In Section II the use case is presented with details of its construction. In Section III we present our process of data collection and subsequent analysis to include statistical exploration and our machine learning approach. We then present the results of our analysis in Section IV followed by conclusions and future direction in Section V.

II. Robot Arm Force-Seeking Application

A. General Construction and Operation

A robotic force-seeking apparatus is constructed using a Universal Robots UR-3 collaborative robot. As illustrated in Fig. 1, the robot is fitted with a six degrees-of-freedom (DOF) force-torque sensor (FTS) followed by a probe. The robot is programmed to apply a downward force, $F(t)$, in the z direction until a force exceeding a threshold, F_t, is reported to the controller. The robot encounters the force threshold through a fixed plunger-spring assembly. The force in the spring is governed by the equation $F(t) = kl$ where k is the spring constant and l is the spring deflection. The robot will push the spring downward repeatedly for the duration of 30 minutes. Plunger movement is limited by a hard stop which will reset the height of the robot arm. A photograph of the force-seeking apparatus is shown in Fig. 2. The illuminated spheres shown in the photo are infrared markers used by the remote observer to track the position of the probe.

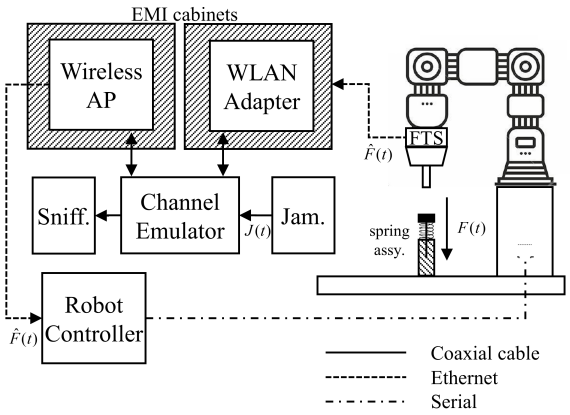


Fig. 1. Robot force-seeking spring system with controlled wireless channel emulation and interference injection.

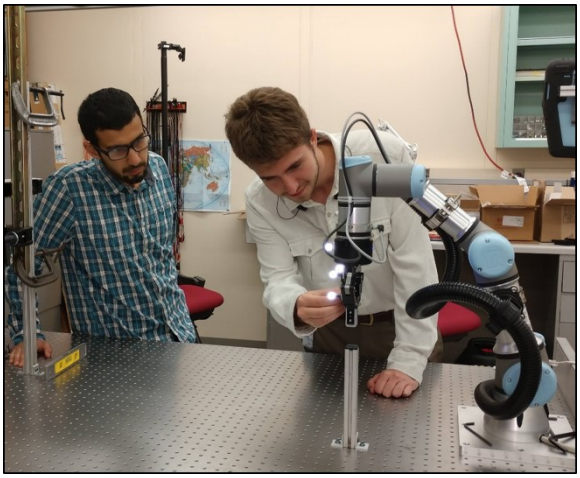


Fig. 2. A photograph of the robot force-seeking experiment shows the robotic arm, the spring-based plunger, and the visual markers used for position tracking.

B. Components

Referring again to Fig. 1, the system is composed of the following components:

Robot The robotic arm applies a downward force along the z-axis to the plunger-spring assembly. The robot is a 6-DOF rigid body manipulator in which all joints have a full 360 degrees of motion. For the experiment, the robot is configured such that it would replicate the action of a robot applying a force to push a small part into place within an automotive assembly work-cell [15]. The robot is mounted on a motionless optics table in which mechanical vibration is dampened.

Robot Controller The robot controller (RC) provides the

Candell, Richard; Montgomery, Karl; Hany, Mohamed; Liu, Yongkang; Foufou, Sebti.
"Wireless Interference Estimation Using Machine Learning in a Robotic Force-Seeking Scenario."
Paper presented at 28th International Symposium on Industrial Electronics (ISIE), Vancouver, Canada. June 12, 2019 - June 14, 2019.

motion control function of all joints on the robot. The RC is responsible for controlling motion while searching for a force feedback signal.

Force Torque Sensor The force torque sensor (FTS) provides continuous force and torque readings at a rate of 125 Hz. Readings from the FTS include force measurements in Newtons along the three Cartesian axes, x, y, and z, and three torque readings in Newton-meters (N-m) about each axis. The FTS is designed to communicate with the RC through an Ethernet connection.

Robot End-effector The robot end-effector (REEF) is a rigid body probe attached to the end of the robot arm just after the FTS. The REEF is used to make contact with the plunger-spring assembly.

Wireless Components The wireless Ethernet adapter (WEA) replaces the Ethernet connection between the FTS and the RC with a Wi-Fi connection. The adapter supports the IEEE 802.11 b, g, n, and ac modes. The WEA connects to the RC through a wireless access point (WAP).

Jammer The jammer provides the source of interference, J, which is directly injected into the wireless channel. For simplicity, interference is injected as non-modulated additive white Gaussian noise (AWGN). The power of J at each receiver is determined by its distance to the jammer.

Channel Emulator The channel emulator (CE) provides the capability to control the electromagnetic channel between the WEA and the WAP. The CE supports frequencies between 1 GHz and 6 Ghz and has an instant bandwidth of 250 MHz. It also supports a channel impulse response of 13 taps with a minimum time resolution of 4 ns making the replication of close-quarter multi-path reflections possible. As shown in Fig. 1, all wireless devices are connected to the CE.

Electromagnetic Interference Cabinets The electromagnetic interference (EMI) cabinets provide isolation between devices such that communication between devices does not occur through radiated leakage.

Wireless Sniffer A wireless sniffer (WS) is used to monitor wireless traffic during operation. The sniffer is connected to a laptop computer running Wireshark, and packet logs are used for offline analysis of network events.

Vision Tracking System An OptiTrack VS120 Trio is used as the vision-based tracking system (VTS) to produce accurate ground truth measurements of the probe position. Position estimates along the z-axis are captured at the maximum video frame rate of 120 frames per second. Each estimate includes time and position.

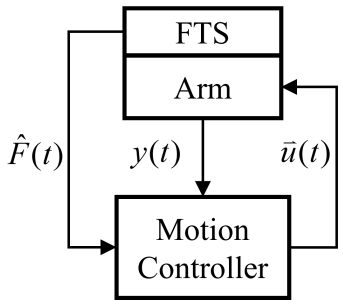


Fig. 3. Feedback signal flow model of the force-seeking controller

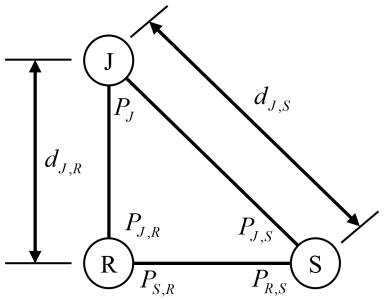


Fig. 4. RF emulation scenario design of the robotic force-seeking scenario.

C. Robot Arm Motion Control

A diagram of the control system for the robotic manipulator is shown in Fig. 3. The UR-3 is constructed of the manipulator assembly and the RC assembly. The internal construction of the robot arm is irrelevant for this experiment, but it is assumed that the arm produces encoder positions $y(t)$ for each joint. It is also assumed that the robot arm accepts actuation signals $\vec{u}(t)$ from the motor drives located in the RC. Both $y(t)$ and $\vec{u}(t)$ are conveyed through wired connections. The force sensor signal $\hat{F}(t)$ is produced by the FTS and is conveyed via an IEEE 802.11 wireless connection. The RC is programmed to move a probe connected to the end of the manipulator downward along a linear path until a force of at least 5 N is detected. The RC will not move the arm during the force-seeking operation unless it receives an FTS signal; therefore, the duration and continuity of the movement of the arm will be impacted by unreliable communication between the FTS and the RC.

D. RF Emulation Scenario

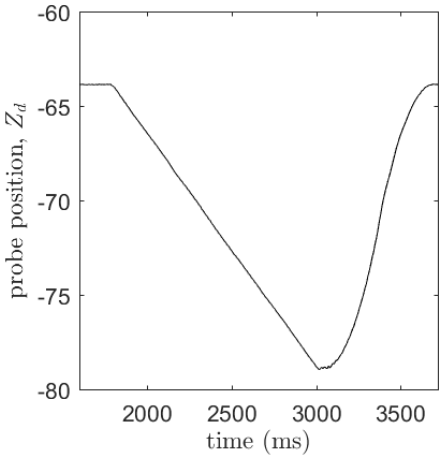
The CE is programmed using a graphical user interface in which the wireless scenario is modeled. Scenarios are composed of radios, platforms, and links. Platforms represent

Candell, Richard; Montgomery, Karl; Hany, Mohamed; Liu, Yongkang; Foufou, Sebti.
"Wireless Interference Estimation Using Machine Learning in a Robotic Force-Seeking Scenario."
Paper presented at 28th International Symposium on Industrial Electronics (ISIE), Vancouver, Canada. June 12, 2019 - June 14, 2019.

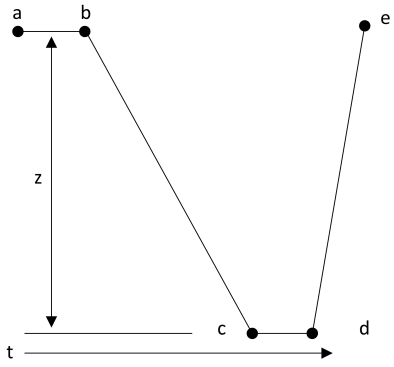
the physical machine on which a radio may be deployed. Platforms may be mobile or stationary, ground-based or aerial. Radios are assigned to platforms, and each radio is associated to a physical port on the emulator. Links are representations of the physical connections between radios. Each link has an associated path loss and multi-path representation. Path loss is implemented according to Friis equation [16] simplified as $P_r = P_t + C - 10\gamma \log_{10}(d)$, where P_r is the received power, P_t is the transmitted power, C is a characteristic constant representing characteristics of the channel and electronics, γ is the path loss exponent, and d is the distance between transmitter and receiver. For simplicity, we assume that path loss occurs in accordance with the square of the distance ($\gamma = 2$); however, in practice, the path loss exponent is usually greater, causing a more rapid loss of signal power over the same distance [17]. Since the focus of this work is to infer signal quality from ground truth measurements, the path loss exponent is inconsequential to our analysis.

Shown in Fig. 4 is the general scenario for the wireless communication system employed for the force feedback control system. In the figure, there are three nodes, a wireless router (R), a wireless station (S), and a jammer (J). The router and station transmit with nominal power that is dependent upon the 802.11 protocol. The jammer transmits with constant power, and its impact on the scenario depends on its position relative to the other nodes. The distance between J and R is denoted by $d_{J,R}$, and the distance between the J and S is denoted by $d_{J,S}$. The resulting signal-to-interference power ratio (SIR) for the router is defined in decibels as $SIR_{J,R} = P_{S,R} - P_{J,R}$ which is the power received by the router of the station signal divided by the power of interference experienced at the router. Similarly, the SIR experienced at the station is defined as $SIR_{J,S} = P_{R,S} - P_{J,S}$ which is the received signal power of the router at the station divided by the interference power experienced at the station.

For each experiment, the location of the J is adjusted to produce a desired SIR. Each time the location of J is changed, the robot is allowed to operate for a period of 30 minutes. This included periods of inaction by the robot when the SIR prohibits movement of the arm. The SIR setting was validated for each run using a real-time spectrum analyzer connected directly to the emulator.



(a) sample time series of probe position



(b) feature extraction model

Fig. 5. A time-series sample (a) of a single iteration of the measured z-axis probe position and (b) the corresponding model for feature extraction.

III. Data Analysis

The data analysis process for the experiment is divided into four parts: raw data collection, data cleaning and feature extraction, training, and the operation of the SIR estimation. The raw data was produced as an output of the VTS as a time series of z-axis position. Feature extraction was conducted in MATLAB by following the time series and extracting or calculating features for each iteration. Once features were extracted, a statistical analysis of the features was conducted to determine the variability of the features as a function of SIR. Statistical analyses included visual inspection of histograms of each factor and an inspection of the correlation coefficients over the range of SIRs. A discussion of the statistical results is provided in IV-A which demonstrates suitability of the use of position measurements for machine learning. Training of a machine learning algorithm followed. The machine learning algorithm was programmed in Python using the Sci-kit Learn library [18].

Candell, Richard; Montgomery, Karl; Hany, Mohamed; Liu, Yongkang; Foufou, Sebti.
"Wireless Interference Estimation Using Machine Learning in a Robotic Force-Seeking Scenario."
Paper presented at 28th International Symposium on Industrial Electronics (ISIE), Vancouver, Canada. June 12, 2019 - June 14, 2019.

A. Feature Extraction

Feature extraction begins with a time series of position of the probe through successive iterations of the plunger applying force to the spring and then returning to its home position. A sample time series of the z-axis position of the probe is as shown in Fig. 5a. Rather than using the time series directly, a more convenient and practical solution is to extract features that represent aspects that may be useful for analysis and machine learning algorithms. This reduces the number of learning dimensions and usually improves computation efficiency. The selected features are illustrated in Fig. 5b. Shown in the model, the probe begins at its home position, **a**. It will not begin its downward motion until it receives sufficient FTS readings. Marker **b** indicates the beginning of the probes descent. Marker **c** represents that point in which the probe descends below a predetermined threshold, and marker **d** represents the position in which the probe begins its return ascent to the home position. Finally the probe returns to the home position as indicated by marker **e**. Therefore, the extracted features of each successive iteration is defined as follows:

Feature Z_d The length of the probe's descent measured in millimeters,

Feature Δt_{ab} The duration in seconds the the robot waits before moving the probe along its descent,

Feature Δt_{bc} The duration in seconds of the time that the robot takes to move the probe beyond the threshold, Z_{th}, of -77 mm,

Feature Δt_{cd} The duration in which the probe dwells below Z_{th} and the speed of the probe remains under 0.15 mm/sec,

Feature Δt_{ae} The duration of the full iteration as measured from the home position, **a**, to the next home position, **e**.

B. Statistical Analysis

Each factor was visually examined to assess its variability as a function of the SIR. In order to predict the SIR given a set of measurements of the dynamics of the physical system, sufficient variability is needed. This assessment was performed visually using histograms as a basis for comparison. The factors Z_d and Δt_{bc} were used for examination of the data using histograms.

In addition, it would be helpful to show that the factors are uncorrelated as a function of SIR demonstrating a further level of confidence that each factor will be useful to a machine learning algorithm. This assessment was accomplished by computing the correlation coefficient matrix of the extracted factors as defined by the Pearson product-moment method [19]. The correlation coefficient matrix is a covariance matrix that is normalized by the product of the standard deviations of two factors being compared according to $\rho_{X,Y} = cov(X,Y)/(\sigma_X \sigma_Y)$. Since each factor correlates exactly with itself, a correlation matrix should have values of 1 along the diagonal. Other elements of the matrix will take on values between -1 and 1. A visual inspection of the coefficient matrices will show how strongly selected factors vary together. Correlation can be viewed as a function of SIR to verify that factors are independently applicable to a learning algorithm. The objective of factor selection is, therefore, to choose factors that are highly uncorrelated and yet still vary appreciably [20].

C. Machine Learning

In order to learn the SIR level from observing the various features, we leverage the random forest model [21]. Random forest is an ensemble of decision trees with random feature selection which can be used for classification or regression based on the predicted output space. Deploying random forest in machine learning has been successful in various applications such as [22]–[24]. Its main advantages are that it is stable, fast to compute, and insusceptible to over-fitting.

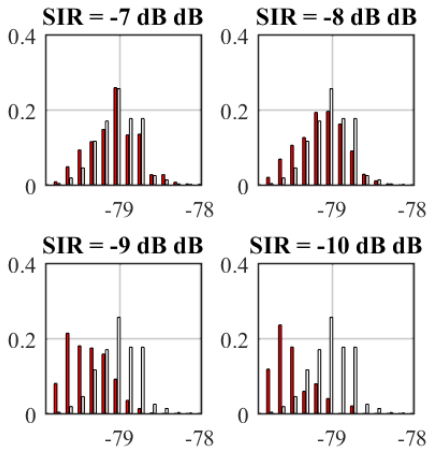
In this work, we deploy the random forest model for SIR regression using the five features defined in III-A. These features are evaluated for each iteration of the probe movement. We define a data segment which is composed of a number of successive iterations and we denote the segment size by M. As a result, we use the random forest regression model to get an input vector of size $5M$ and regression output of the corresponding SIR value. The random forest is selected because it is computationally efficient with high-dimensional data and it is robust for outliers and data non-linearity.

We start by training the random forest regression model by taking a fixed number of segments for each SIR labelled data. We denote the size of the training set for each SIR level by T. The rest of the measurements are used for testing. In general, the proposed machine learning approach will deploy a sliding window approach of size M to collect the features of the force-seeking use case to estimate the current level of SIR at various nodes of the wireless network.

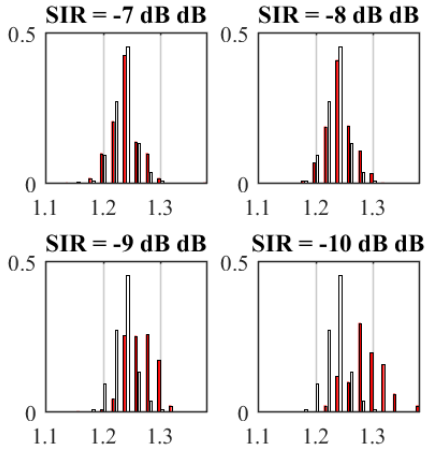
IV. Results

The results in this section are presented from an experimental run in which the jammer **J** interferes with the router, **R**, while communication is conducted using a mixed mode of IEEE 802.11 b and g [14]. Analyses using histograms and covariance are presented in Section IV-A followed by results of the machine learning application in Section IV-B.

Candell, Richard; Montgomery, Karl; Hany, Mohamed; Liu, Yongkang; Foufou, Sebti.
"Wireless Interference Estimation Using Machine Learning in a Robotic Force-Seeking Scenario."
Paper presented at 28th International Symposium on Industrial Electronics (ISIE), Vancouver, Canada. June 12, 2019 - June 14, 2019.



(a) z-axis position



(b) Plunge delay Δt_{bc}

Fig. 6. Variations in probability distributions of the z-axis position (a) and the plunge delay (b) indicate that machine learning may be effective in inferring information about the underlying communication channel. In the figure, the baseline case of infinite SIR is depicted as a histogram with white bars, and the experimental case is depicted as a histogram with red bars.

A. Statistical Analysis

1) Analysis of Factors Using Histograms: The results of the histogram analyses for the z-axis position of the probe and the probe descent delay are shown in Fig. 6a and Fig. 6b, respectively. The expectation of the histogram analysis was that Z_d and the ΔT_{bc} would exhibit appreciable variation that may be observed through a visual inspection. This was indeed the case. Referring to Fig. 6a, a visual inspection reveals that the minimum z-axis position for each iteration skews to lower positions for lower SIR values and higher positions for higher SIR values. This implies that the controller algorithm responds faster to force sensor readings at higher SIR values than lower values. Similarly, by observing the plunge delay, Δt_{bc}, the controller takes more time to respond at lower SIR values than at higher values. This behavior is exemplified by the probability skew shown in the histograms.

TABLE I
CORRELATION COEFFICIENTS FOR $-9\ dB$ SIR

	Δt_{ab}	Δt_{bc}	Z_d	Δt_{cd}	Δt_{ae}
Δt_{ab}	1	0.04	0.04	0	0.56
Δt_{bc}	0.04	1	-0.96	-0.08	0.18
Z_d	0.04	-0.96	1	0.03	-0.01
Δt_{cd}	0	-0.08	0.03	1	0
Δt_{ae}	0.56	0.18	-0.01	0	1

TABLE II
CORRELATION COEFFICIENTS FOR $-8\ dB$ SIR

	Δt_{ab}	Δt_{bc}	Z_d	Δt_{cd}	Δt_{ae}
Δt_{ab}	1	0.01	-0.05	-0.06	0.58
Δt_{bc}	0.01	1	-0.99	-0.13	0.1
Z_d	-0.05	-0.99	1	0.07	-0.1
Δt_{cd}	-0.06	-0.13	0.07	1	-0.03
Δt_{ae}	0.58	0.1	-0.1	-0.03	1

TABLE III
CORRELATION COEFFICIENTS FOR $-7\ dB$ SIR

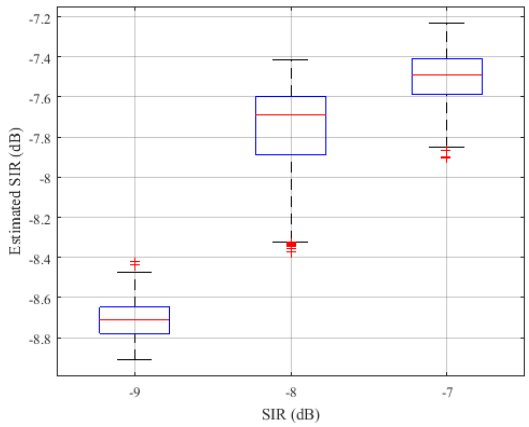
	Δt_{ab}	Δt_{bc}	Z_d	Δt_{cd}	Δt_{ae}
Δt_{ab}	1	-0.05	-0.04	0.09	0.01
Δt_{bc}	-0.05	1	-0.93	-0.17	0.05
Z_d	-0.04	-0.93	1	0.08	-0.05
Δt_{cd}	0.09	-0.17	0.08	1	-0.02
Δt_{ae}	0.01	0.05	-0.05	-0.02	1

2) Factor Correlation Coefficient Analysis: Correlation coefficients were calculated for each of the five factors defined in Section III-A and correlation coefficients matrices were produced for each of the SIR values used. The correlation coefficient matrices for SIR values of -9, -8, and -7 are shown in Tables I-III, respectively. Inspection of the correlation coefficient tables indicate that the factors are mostly uncorrelated across SIR values except for the clear correlation between plunge delay and plunge depth. Low correlation values demonstrate a necessary but not sufficient condition for the independent applicability factors to machine learning. If desired, either Δt_{bc} or Z_d could be omitted as they are strongly correlated and therefore provide redundant information.

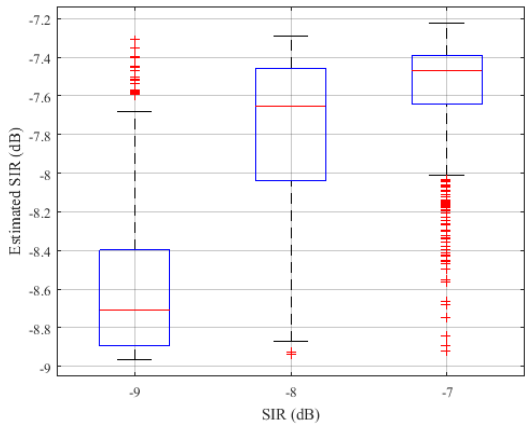
B. Machine Learning Results

We deploy the proposed machine learning approach to three values of the SIR, -9, -8, and -7 dB. We start by showing the output of the random forest regression model for two values

Candell, Richard; Montgomery, Karl; Hany, Mohamed; Liu, Yongkang; Foufou, Sebti.
"Wireless Interference Estimation Using Machine Learning in a Robotic Force-Seeking Scenario."
Paper presented at 28th International Symposium on Industrial Electronics (ISIE), Vancouver, Canada. June 12, 2019 - June 14, 2019.



(a) $M = 100$



(b) $M = 1$

Fig. 7. Predicted SIR versus actual SIR for the cases of (a) $M = 100$ and (b) $M = 1$. The box plots show the median value while the bottom and top edges of the box indicate the 25th and 75th percentiles. Statistical outliers are shown as red $+$ signs.

of the segment size M. We set the training set size $T = 200$ for each SIR value. We use the random forest model with a number of estimators of 500 and a tree depth of 5. In Fig. 7, we present the box plots of the predicted SIRs against the correct value of the corresponding SIR for $M = 100$ and $M = 1$. Generally, increasing the value of M increases the acquisition time for the input data for the random forest model while enhancing the performance of the algorithm. By setting $M = 1$, we notice that the predicted values of SIR are widely spread around the median and a large number of outliers exists. However, by increasing M, we have much less variations in the predicted SIRs and a smaller number of outliers.

In Fig. 8, we present the two criteria for measuring the

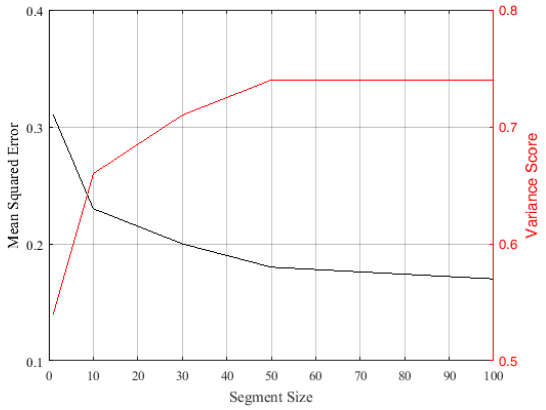


Fig. 8. The performance of the random forest regression model against M.

performance of the proposed SIR estimation algorithm. We show the performance against the segment size M. The first criterion is the mean squared error where the mean of the squared error between the estimated SIR and the actual SIR values is calculated. The second criterion is the variance score which is a statistical measure of how close the data are to the fitted regression line. We use the r-squared variance score that is defined as the ratio between the total variance explained by model and total variance of the data [25]. In this figure the improvement in the performance against the segment size is demonstrated.

V. Conclusion

In this paper we have presented a practical use case of a wireless force-torque feedback control system that could be deployed in a manufacturing assembly system such as a pick-and-place or assembly operation. A 6-DOF force sensor was connected to a robot controller tasked with moving a probe along a linear path until an opposing force exceeding 5 N was detected. We demonstrated that the reliability of the wireless communication system directly impacts the repeatability performance of the physical system. We also demonstrated that the quality of the underlying wireless channel may be inferred by observing the position of the probe along a single spatial dimension and applying machine learning to predict the signal-to-interference ratio. Our findings provide motivation for applying machine learning to larger more complex systems with high degrees of freedom. Future work will extend to the inclusion of more descriptive factors, the addition of network information such as the wireless protocol mode, and the addition of a larger number of variables tracked by many remote observers. Experimentation with neural networks

Candell, Richard; Montgomery, Karl; Hany, Mohamed; Liu, Yongkang; Foufou, Sebti.
"Wireless Interference Estimation Using Machine Learning in a Robotic Force-Seeking Scenario."
Paper presented at 28th International Symposium on Industrial Electronics (ISIE), Vancouver, Canada. June 12, 2019 - June 14, 2019.

and deep learning to improve prediction accuracy and better generalization will be of great values to wireless operations in factories. Finally, the applications of online machine learning techniques to this and other use cases could provide significant benefits to the manufacturing community.

DISCLAIMER

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

[1] R. Candell, M. Hany, K. B. Lee, Y. Liu, J. Quimby, and K. Remley, "Guide to industrial wireless systems deployments," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep., Apr 2018. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/ams/NIST.AMS.300-4.pdf

[2] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial Internet of Things: Challenges, Opportunities, and Directions," *IEEE Transactions on Industrial Informatics*, 2018.

[3] L. L. Bello, J. Åkerberg, M. Gidlund, and E. Uhlemann, "Guest Editorial Special Section on New Perspectives on Wireless Communications in Automation: From Industrial Monitoring and Control to Cyber-Physical Systems," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1393–1397, 2017.

[4] Z. Pang, M. Luvisotto, and D. Dzung, "Wireless High-Performance Communications: The Challenges and Opportunities of a New Target," *IEEE Industrial Electronics Magazine*, vol. 11, no. 3, pp. 20–25, 2017.

[5] R. Candell, "Industrial wireless systems workshop proceedings," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep., May 2017. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8174.pdf

[6] C. Lu, A. Saifullah, B. Li, M. Sha, H. Gonzalez, D. Gunatilaka, C. Wu, L. Nie, and Y. Chen, "Real-Time Wireless Sensor-Actuator Networks for Industrial Cyber-Physical Systems," *Proceedings of the IEEE*, vol. 104, no. 5, pp. 1013–1024, May 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7348717/

[7] D. Kim, Y. Won, Y. Eun, and K.-J. Park, "W-Simplex: Resilient network and control co-design under wireless channel uncertainty in cyber-physical systems," in *2017 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, Aug 2017, pp. 49–54. [Online]. Available: http://ieeexplore.ieee.org/document/8062439/

[8] T. M. Chiwewe, C. F. Mbuya, and G. P. Hancke, "Using Cognitive Radio for Interference-Resistant Industrial Wireless Sensor Networks: An Overview," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 6, pp. 1466–1481, Dec 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7299315/

[9] N. Bhushan, Junyi Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, Feb 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6736747/

[10] R. D. Gomes, D. V. Queiroz, A. C. Lima Filho, I. E. Fonseca, and M. S. Alencar, "Real-time link quality estimation for industrial wireless sensor networks using dedicated nodes," *Ad Hoc Networks*, vol. 59, pp. 116–133, May 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1570870517300434

[11] N. Baccour, A. Koubâa, L. Mottola, M. A. Zúñiga, H. Youssef, C. A. Boano, and M. Alves, "Radio link quality estimation in wireless sensor networks," *ACM Transactions on Sensor Networks*, vol. 8, no. 4, pp. 1–33, Sep 2012. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2240116.2240123

[12] U. Wetzker, I. Splitt, M. Zimmerling, C. A. Boano, and K. Romer, "Troubleshooting Wireless Coexistence Problems in the Industrial Internet of Things," in *Proceedings - 19th IEEE International Conference on Computational Science and Engineering, 14th IEEE International Conference on Embedded and Ubiquitous Computing and 15th International Symposium on Distributed Computing and Applications to Business, Engi*, 2017.

[13] G. H. Koepke, W. F. Young, J. M. Ladbury, and J. B. Coder, "Interference and Coexistence of Wireless Systems in Critical Infrastructure," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep., Jul 2015. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.1885.pdf

[14] IEEE Standards Association, *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications; Amendment 4: Enhancements for Very High Throughput for Operation in Bands below 6 GHz*, 2013.

[15] M. L. T. Cossio, L. F. Giesen, G. Araya, M. L. S. Pérez-Cotapos, R. L. VERGARA, M. Manca, R. A. Tohme, S. D. Holmberg, T. Bressmann, D. R. Lirio, J. S. Román, R. G. Solís, S. Thakur, S. N. Rao, E. L. Modelado, A. D. E. La, C. Durante, U. N. A. Tradición, M. En, E. L. Espejo, D. E. L. A. S. Fuentes, U. A. D. Yucatán, C. M. Lenin, L. F. Cian, M. J. Douglas, L. Plata, and F. Héritier, *Springer Handbook of Robotics*, B. Siciliano and O. Khatib, Eds. Springer International Publishing, 2016. [Online]. Available: http://link.springer.com/10.1007/978-3-319-32552-1

[16] J. A. Shaw, "Radiometry and the Friis transmission equation," *American Journal of Physics*, vol. 81, no. 1, pp. 33–37, Jan 2013. [Online]. Available: http://aapt.scitation.org/doi/10.1119/1.4755780

[17] R. Candell, C. Remley, J. Quimby, D. Novotny, A. Curtin, P. Papazian, G. Koepke, J. Diener, and M. Kashef, "Industrial wireless systems: Radio propagation measurements," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep., 2017. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.1951.pdf

[18] "scikit-learn: machine learning in Python." [Online]. Available: http://scikit-learn.org/stable/

[19] K. Yeager, "LibGuides: SPSS Tutorials: Pearson Correlation." [Online]. Available: https://libguides.library.kent.edu/SPSS/PearsonCorr

[20] J. Lee Rodgers and W. Alan Nice Wander, "Thirteen ways to look at the correlation coefficient," *American Statistician*, pp. 59–66, 1988.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[22] X. Zhen, Z. Wang, M. Yu, and S. Li, "Supervised descriptor learning for multi-output regression," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1211–1218.

[23] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*, June 2011, pp. 1297–1304.

[24] X. Zhen, Z. Wang, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Direct estimation of cardiac bi-ventricular volumes with regression forests," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds. Cham: Springer International Publishing, 2014, pp. 586–593.

[25] S. Deb, "A novel robust r-squared measure and its applications in linear regression," in *Computational Intelligence in Information Systems*, S. Phon-Amnuaisuk, T.-W. Au, and S. Omar, Eds. Cham: Springer International Publishing, 2017, pp. 131–142.

Candell, Richard; Montgomery, Karl; Hany, Mohamed; Liu, Yongkang; Foufou, Sebti.
"Wireless Interference Estimation Using Machine Learning in a Robotic Force-Seeking Scenario."
Paper presented at 28th International Symposium on Industrial Electronics (ISIE), Vancouver, Canada. June 12, 2019 - June 14, 2019.

Towards Standard Exoskeleton Test Methods for Load Handling

Roger Bostelman, Ya-Shian Li-Baboud, Ann Virts, Soocheol Yoon, Mili Shah

***Abstract*— Exoskeletons are now being marketed by several manufacturers and yet there are currently no standard test methods to help match exoskeletons[1] to desired tasks. The National Institute of Standards and Technology (NIST) has been a key contributor to the formation of a new ASTM F48 standards committee on exoskeletons and has a project to research measurement science and test methods in support of exoskeleton standards. This paper describes the NIST exoskeleton project efforts for one of several ongoing research areas that target typical industrial tasks - i.e., load handling. The paper will describe the design of the NIST Position and Load Test Apparatus for Exoskeletons (PoLoTAE), a reconfigurable testbed and its use within a NIST human subject's study. Experimental results from the first load handling test are described and future tests are outlined.**

I. Introduction

Exoskeletons can be passive, with only springs and/or counterweight augmentation of human motion, or active, with motor augmentation - sometimes called wearable robots, or a combination of passive and active. In the early 2000's, the U.S. Defense Advanced Research Project Agency (DARPA) began development and demonstration of "… critical technologies such as power, control, and actuation that will lead to a self-powered external structure to enable a Soldier to effortlessly carry over 45 kg (100 lbs) of additional weight [1]." Exoskeletons were recognized as potentially beneficial to many other domains: as of December 2016, [2] identified 58 commercial and/or non-profit organizations that have developed or were developing exoskeletons or wearable robotics. Today, there are roughly 80 commercial systems on the market.

Several events laying the foundation for standards have occurred since the DARPA developments, including a Round Table and several Technical Interchange Meetings, fostered by U.S. government organizations and with invitees from international government, industry, and academia [3]. One common theme that was discussed was the need for standard test methods for exoskeletons. To this end, an exploratory project within the National Institute of Standards and Technology (NIST) Engineering Laboratory (EL), began in September 2017 and cumulated into the ASTM International Committee F48 Exoskeletons and Exosuits (ASTM F48) [4]. Six F48 subcommittees were established: F48.01 Design and Manufacturing, F48.02 Human Factors and Ergonomics, F48.03 Task Performance and Environmental Considerations, F48.04 Maintenance and Disposal, F48.05 Security and Information Technology, and F48.91 Terminology. An exoskeleton is currently defined by ASTM F48.91 as a "wearable device that augments, enables, assists, and/or enhances physical activity." Two notes support the definition with 1) "an exoskeleton may include rigid and/or soft components (see exosuit)" and 2) "physical activity may be static or dynamic".

The NIST EL project also includes developments in two areas that have continued into a longer-term project beginning in fiscal year 2019: 1) measurement science towards development of new methods to measure the exoskeleton fit and movement of the exoskeleton to the user, and 2) the impact that wearing an exoskeleton has on the performance of users executing tasks that are representative of activities in industrial settings. Based on the United States Bureau of Labor statistics, in 2014 there were about 12 188 300 manufacturing jobs in the United States [5]. Approximately 5 086 905 of manufacturing employees belong to small and medium enterprises according to the US Census Bureau for 2012 [6]. With so many manufacturing jobs performed, including nearly half in the small and medium size organizations, even a small percentage of these jobs could benefit from using exoskeletons if they are safe and effective. As a result, exoskeleton standards are necessary.

Initial NIST EL efforts to support exoskeleton standards include developing test methods from previous research areas such as response robots and autonomous industrial vehicles [7]. A NIST study, which began in June 2018 [8], is researching the two EL project areas described above. The results of the study will inform future test method development at NIST and at other organizations, all under the purview of ASTM F48. Area 1 has been completed for exoskeleton fit to the leg and analysis is underway. Within area 2, the focus of this paper, a reconfigurable Position and Load Test Apparatus for Exoskeletons (PoLoTAE) [9] was designed. Six test methods were designed to target industrial tasks where the first load positioning task will be the focus experiment and findings described in this paper. The outcome of the load handling tasks is also expected to provide support for development of ASTM

R. Bostelman is with the National Institute of Standards and Technology, Gaithersburg, MD 20899, USA (phone: 301-975-3426; e-mail: roger.bostelman@nist.gov).

Ya-Shian Li-Baboud is with the National Institute of Standards and Technology, Gaithersburg, MD 20899, USA (e-mail: yashian.libaboud @nist.gov).

Ann Virts is with the National Institute of Standards and Technology, Gaithersburg, MD 20899, USA (e-mail: ann.virts@nist.gov).

Soocheol Yoon is with the National Institute of Standards and Technology, Gaithersburg, MD 20899, USA (e-mail: soocheol.yoon @nist.gov).

Mili Shah is with The Cooper Union for the Advancement of Science and Art, Department of Mathematics, New York, NY 10003 (e-mail: mili@cooper.edu). Research for Shah is performed under grant 70NANB17H251 from U.S. Department of Commerce, National Institute of Standards and Technology.

[1] Disclaimer: Commercial equipment, software, and materials are identified in order to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials, equipment, or software are necessarily the best available for the purpose.

Bostelman, Roger; Li-Baboud, YaShian; Virts, Ann; Yoon, Soocheol; Shah, Mili.
"Towards Standard Exoskeleton Test Methods for Load Handling."
Paper presented at WearRAcon 19, Scottsdale, AZ, United States. March 26, 2019 - March 28, 2019.

subcommittee F48.03 work item WK65295 - Load Handling When Using an Exoskeleton.

II. Test Method Design

Repeatable test methods can help exoskeleton manufacturers and users highlight capabilities of their systems, compare exoskeletons to their motion tasks, show design flaws or enhancements, and help with procurement requirements. Ideally, these test methods are not only repeatable, but also standardized, such that both manufacturers and users can simply select a document that describes how to perform the test method no matter which exoskeleton they make or use. Although NIST has been focusing on industrial test methods, there is overlap with military, medical, response, and perhaps commercial exoskeletons. For example, military logistics personnel carry relatively heavy loads and, for example, mount wheels on military equipment. Example cross-industry/medical use of exoskeletons may be for nurses and orderly staff to pick-up and maneuver patients. Movement of an elderly person at home with the help of an exoskeleton device would not be considered a medical application if the device is not prescribed by a medical doctor.

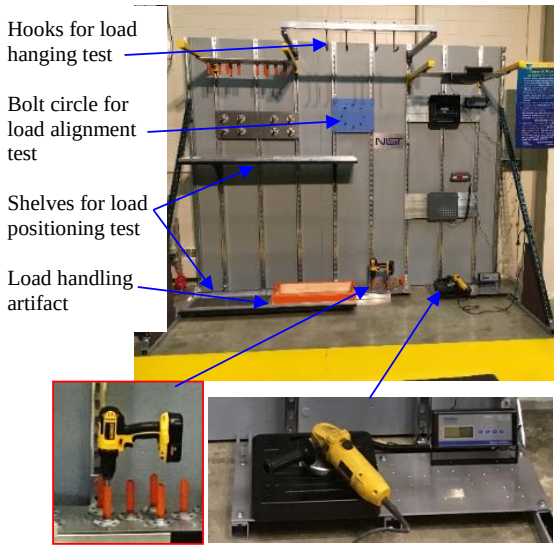


Figure 1. PoLoTAE showing inserts of the force plate with tool (top) and drill with bit in a pipe for peg-in-hole tests.

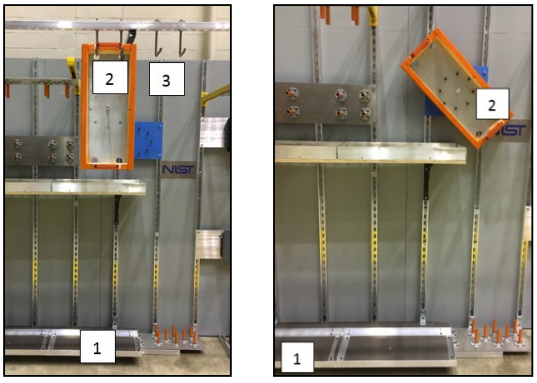
In addition to supporting the development of ASTM F48, which includes all of the above applications, NIST began designing test methods that are representative of industrial tasks including: applying forces (e.g., grinding, sawing, drilling, pushing, etc.), inserting pegs (e.g., screw, drill bit, etc.) into holes, non-contact alignment (e.g., laser pointing, siting, etc.), and load handling (e.g., holding a load and aligning to a fixture, hanging a load on hooks, and placing/positioning loads). All of these tests are planned for the experiments within the NIST study. In this paper, we describe the study of the first of the six planned tasks – i.e., load positioning.

Towards a standardized approach to a replicable apparatus that can be used by the exoskeleton manufacturers, users, and researchers, NIST designed the PoLoTAE and will provide the PoLoTAE design in a future NIST internal report [8]. The PoLoTAE, shown in Figure 1, was designed to be reconfigurable, minimal cost, and expandable to larger dimensions or to focus more specifically on the required exoskeleton tasks. Variability in generic loads/tools, positioning heights, defined motion spaces, etc. can also be added to the PoLoTAE to help exoskeleton designers fit the tasks to the potentially wide variety of exoskeleton wearers.

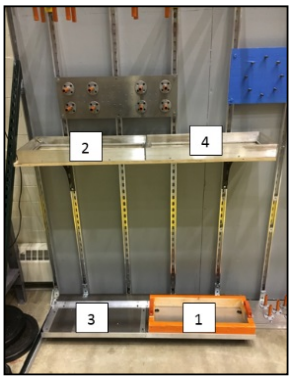
Loads can vary dramatically depending on the application. However, load handling includes not only picking up but also placing loads, sometimes with relative precision (e.g., installing a wheel on a vehicle axle bolt- circle). Ideally, a single, replicable artifact (e.g., the load artifact shown in Figure 1) is used for picking, placing, aligning, and hanging loads.

Figure 2 a, b, and c show the series of load handling tasks that are planned for experimentation on at least 30 subjects each. Figure 2 a and b show load hanging and load alignment tests planned for future experiments within the study and will not be further explained in this paper. Figure 2 c shows the load positioning test design setup which is the focus of the remainder of this paper.

Figure 2 d shows the 6.8 kg (15 lbs) load artifact designed for use in all three load handling tests. For load positioning, the artifact base was approximately 3 mm smaller on all four sides than the tray-surround on each shelf. The surround ensures that the subject being tested must not only place the artifact in the correct location but must also position and seat the artifact correctly within the surround. Handles or other grasp methods can also be incorporated into the artifact.



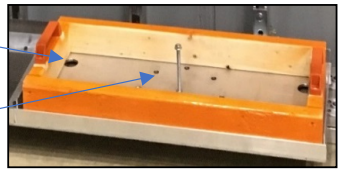
a b



c

Holes for load hanging – 2 on each end

Holes for load alignment



d

Figure 2. PoLoTAE a) load hanging, b) load alignment, and c) load positioning test designs setup and numbered order of subject movements planned within the NIST study. d) Artifact used for all three load handling tasks.

III. Load Positioning Experiment

NIST currently owns a full-body, passively-mechanized exoskeleton which uses springs internal to the frame. This exoskeleton will allow for the initial development of study procedures with the expectation that additional models will be procured or borrowed to validate and expand the tests on a variety of exoskeletons. The test methods described below are designed to exercise all aspects of a full-body exoskeleton through timed video recording, so that the sub-tasks (e.g., lifting, placing, bending, etc.) can be separated out for individual review.

Subject recruitment included posting brochures requesting NIST federal employee volunteers that fit the profile required of the test and the exoskeleton to be used. In a similar test method development, EL researchers worked with the NIST Statistical Engineering Division to determine the number of tests to perform for statistical significance. The statisticians determined that for 85 % confidence, 30 repetitions were required. At least 30 subjects, each performing one of six tests, requires nearly 200 subjects. Also, the same subject can return for another future test. From a pilot study on a subject having average physicality, 30 repetitions of a task appeared to be a good limit to test fatigue without strain caused by the task motions.

The recruited subjects were:

- At least 18 years of age
- Physically fit to complete the test which relied on the subjects perspective of their own capability.
- Willing to participate for up to 1 ½ hours.
- Able to wear an exoskeleton that weighs approx. 13.7 kg (30.3 lbs).
- Able to perform knee bends, position tools, and apply forces 30 times twice (60 total) using tools (up to the approx. weight of 13.7 kg (30.3 lbs)).
- Able to fit within the exoskeleton manufacturers specification for height 1.5 m to 1.9 m (5’ 0” to 6’ 1”) and weight 49 kg to 102 kg (108 lbs to 225 lbs).

The research study procedure included the following:

- the subject reviewed a training video [11] and signed a consent form,
- the subject wore (one or two) wrist heart rate (HR) monitor(s) and the research team recorded the subject’s HR throughout the test,
- the subject first performed the baseline and then exoskeleton-use tasks, 30 times each, with help from the research team as needed to assist the subject with putting on/taking off the exoskeleton where the subject could stop the test at any time,
- the subject responded to a brief set of survey questions about the test upon its completion.

Repetition 1: The subject stepped one step forward from a start line to the load apparatus (refer to Figure 2c); the load apparatus was picked up from tray 1 (14 cm above the floor) and placed in tray 2 (138 cm above the floor); hands were released; The load apparatus was picked up from tray 2 and placed in tray 3; hands were released and the subject stepped back to the line. Repetition 2: The subject stepped from the line to the load apparatus; the load apparatus was picked up from tray 3 and placed in tray 4; hands were released; The load apparatus was picked up from tray 4 and placed in tray 1; hands were released, and the subject stepped back to the line. Repetition 1 and 2 are repeated 14 more times for a total of 30 repetitions.

IV. Subject Safety Protection

Subject safety - including physical safety as well as social safety - is a priority during the NIST exoskeleton research. All procedures during the study were approved by the NIST Institutional Review Board. The research team strictly followed the Human Subject Research protocol [8] to keep the subject socially safe. In addition, the subject is prevented from physical injury due to instability and raised heart rate through continuous monitoring.

To protect the subject’s personal information, the entire test area is surrounded with curtains. A pre-test survey is conducted to check whether the subject has potential risks (e.g., aches, pains, past broken bones, etc. that may affect their task completion). Many exoskeleton novices were unstable at first and were asked by the team to perform knee bends on soft safety mats to establish their stability. All wearable devices and parts were cleaned with sanitary wipes before and after the test, particularly for the components in contact with the body.

A. Heart Rate Monitoring

HR was monitored throughout the test. HR has been widely used as an indicator for response of heart to exercise

since it is easy to measure and directly related to the autonomic nervous system [12]. The tests require activity and as a result HR was expected to increase during the tests. The American Heart Association [13] method of subtracting the subject age from 220 beats per minute (bpm) or a particular HR that the subject desired was set as the HR limit.

Wrist HR monitors, now widely used [14], were employed to measure and record HR. The monitors use optical sensors to measure HR which are verified to have 80 % to 91 % accuracy depending on make, application, test conductor, and subject response [15][16][19]. The actual accuracy of the HR monitor varies according to the subject's skin type, arm thickness, and shape. In addition, since the tasks require a wide range of motion, errors in HR measurement were primarily due to loss of steady attachment for durations up to 10 s.

To overcome this limitation, the procedure was revised during the experiment to include two steps to provide further corroboration of HR data. First, two HR monitors from different manufacturers were used. Both use optical sensors, but they have different sensor arrangements and straps. When they showed different HR values, the higher one was always accepted. Second, a hand grip monitoring bar, based on electrical conductance, was used to verify that the wrist HRs were working correctly. The bar uses electrodes to measure HR and was used as another technology method to help verify the optical HR monitoring method. Before the test, resting HR was measured, and the highest wrist HR was recorded after the test was verified with the HR bar.

Wrist HR was measured and recorded at 1 s time intervals using BlueTooth to offboard heart rate monitoring devices. Several HRs were marked at specific task repetitions such as prior to beginning, after 15, and after 30 repetitions. The research team frequently asked subjects their condition when their HR was elevated, e.g., above 150 beats per minute (bpm), and determined whether to continue the test, rest, or stop the test.

V. Data Collection System

The data collection system was comprised of four high definition (HD) camcorders mounted on tripods, a high definition multi-viewer, and an HD video capture device with external hard drive. The video was recorded at full HD 1080 pixel resolution. Figure 3 shows the output of the data collection system. Two videos of the output were recorded: the baseline test and the test when the subject wears the exoskeleton.

Figure 3. Snapshot of the data collection system display.

Frame A denotes (top) repetition number, (center) global positioning system time, temperature, date, and humidity, and (left) a timer. For the first load positioning task test and if time were a critical metric, video review could have been used to determine repetition-completion count within a set time. Figure 3 shows the more recent data collection system with repetition counter and timer added for the second task (e.g., peg-in-hole task currently underway). Frame B shows the output of two HR monitors (left and right) as described in section IV. Frames C and D are left and right views of the task being performed by the subject.

Digital photographs of the subject performing various portions of the task were also captured and added to the stored data. For example, the photos show snapshots of:

- the technique that the subject uses to pick-up and place the load artifact (e.g., bend at the hips or bend at the knees),
- the shelf height relative to the subject shoulder height (i.e., potentially critical for short subjects and less critical for tall subjects),
- and the change in the subject form from when first using the exoskeleton to final repetitions after the subject appears to be familiar with the exoskeleton.

VI. Results

A. *Use of heart rate data*

HR was monitored to protect the subject and also used to evaluate the physical demands of the task test on the subject with and without the exoskeleton. It is difficult to conclude whether a subject's HR response can be an indicator of an exoskeleton's effectiveness in augmenting or supporting humans with the physical demands of the task. However, combining HR with other test data obtained including video and a survey, additional information may be derived. Usually HR becomes high at the beginning when wearing an exoskeleton, and later it differs for each subject. For example, if a subject feels uncomfortable, HR gradually increases until the test is finished. Meanwhile, if a subject seems comfortable with using the exoskeleton, HR decreases or remains fairly constant. Even if a subject feels good, his/her HR can be higher than the reference. Thus, rather than the HR value itself, the HR value trend tells us much more about the test and performance.

B. *Data for Load Positioning*

The data set for load positioning included fit, HR, and survey data sets for 33 subjects who completed the Load Positioning task. Each subject completed the tasks in the following order, baseline 30 repetitions (without the exoskeleton), and 30 repetitions with the exoskeleton after a period of returning to a normal, resting heart rate. The uncertainty of the data sets are as follows:

1. The exoskeleton fit based on the anthropometry of the subject was estimated by asking the subject for rough height and waist measurements, which were typically given in inches, while the analysis was done in

Bostelman, Roger; Li-Baboud, YaShian; Virts, Ann; Yoon, Soocheol; Shah, Mili.
"Towards Standard Exoskeleton Test Methods for Load Handling."
Paper presented at WearRAcon 19, Scottsdale, AZ, United States. March 26, 2019 - March 28, 2019.

centimeters. A rough estimate of the size uncertainty is around 3 cm, based on the 2.54 cm per inch conversion. Fit uncertainty is based on comfort of the user and visual verification of exoskeleton alignment to the subject's joints. In particular, the shin and femur adjustments are done based on alignment to the subject's heels, hips and knees, ensuring the exoskeleton is aligned to the center of the sagittal plane, where the inseam of the pants provide a sufficient visual indicator. The arms are based on a few inches above the elbow. The team received training as well as guidance from the exoskeleton manufacturer in fitting the suit to subjects.

2. The HR is estimated using two, and up to three devices, including both the use of optical wrist watches, and an electrical handheld device for corroboration. The HR recorded at the beginning of the task, while the subject is seated, and at the completion time of the task is based on the highest bpm of either of the optical devices. Of the optical devices, one was used predominantly for most of the Load Positioning subjects, and the other was only acquired later in the experiment. The electrical device is used only to verify the optical devices. If the steady heart rate is greater than 10 bpm compared to the optical wristwatch device(s), the electrical device reading was used. Issues that degrade the accuracy of the optical devices include the fit of the wristwatch, sensitivity to perspiration, and other factors that can impact the ability to maintain adequate skin contact for accurate readings. Studies have shown the optical wrist-worn monitors have a concordance correlation of 0.83 to 0.91 at a 95 % confidence interval [16].
3. The survey is based on a series of quantitative categories and free-form answers regarding pre-existing discomforts of the subject and the respective severity, comfort, range of motions, fit, discomfort to the subject caused by the exoskeleton or the task, and the subject's opinion on where the exoskeleton provided support during the task. While the full survey allowed users to enter 33 different regions of the body at 4 severity levels, *slight*, *moderate*, *severe*, and *extreme*, we based the initial analysis on just 10 areas that were indicated by the pool of initial subjects. The areas included head, neck, upper spine, mid-spine, back, lower spine, shoulders, waist, and hips. The free-form answers were also reviewed and categorized to ease analysis. For example, the responses to what the subject liked most about the exoskeleton included: (1) arm support, (2) back support, and (3) leg support. Similarly, the subjects least favorite attributes of the exoskeleton include: (1) heat dissipation, (2) knee motion, (3) range of motion, (4) fit, (5) stability, (6) arm motion/resistance, (7) impact on fine motor movements, and (8) exoskeleton weight. The survey also provided the ratings of perceived discomfort (RPDs) [10] based on Likert scales of 0 (uncomfortable) to 5 (very comfortable) as well as the actual regions of discomfort before, during, and after the test.
4. A quad video of the time and environmental conditions, HR monitors, and the left and right side of subject performing the task was also taken. Our current use of the video is limited to visual assessment of subject's skeletal joint positions. The video has also been used to corroborate with other data sources including HR, RPD, etc. We anticipate expanding the computational analysis by using the videos for quantitative assessment on skeletal joint angle variability.

C. *Analysis method*

An effective test methodology assesses the performance or exertion of the subject while wearing the exoskeleton and without the exoskeleton to complete the load positioning task. The intra-subject independent variable is the exoskeleton. The initial exploratory analysis is based on the Spearman's rank correlation coefficient, which assesses whether there is a monotonic relationship between two variables. Spearman's rank correlation can be used on both discrete ordinal (gender, exoskeleton fit sizes, survey agreement) and continuous (height, waist, heart rate differential) variables.

Various methods have been documented in applying heart rate to capture exercise intensity including the Karvonen method [17] and variants [18], both of which require age information. Because age was not available, the heart rate differential with and without the exoskeleton was used as a metric to capture the physical intensity of the task to the subject:

$$\Delta HR_{LP} = \left(Exo_{end} - Exo_{begin}\right) - \left(Baseline_{end} - Baseline_{begin}\right)$$

Equation 1

The final differential with and without the exoskeleton is recorded using the following equation:

$$\Delta HR_{Final} = \left(Exo_{end}\right) - \left(Baseline_{end}\right)$$

Equation 2

D. *Table of Results*

Tables 1 and 2 show a few trends above 95% level of significance with respect to gender, anthropometry, and fit. Based on [19], the r_{crit} value of the R coefficient, with 30 subjects is 0.306 at the 95 % confidence level. Females and those who indicated prior pains generally had lower heart rates in wearing the exoskeleton relative to the baseline. Shoulder width also indicated a correlation of broader shoulder width to an increase in the subject's physical exertion.

Table 1. Spearman rank correlation to final heart rate differential, ΔHR_{Final}.

	R-Coefficient	P-Value
Gender	-0.36	0.04
Rigid Waist (exo fit)	0.35	0.05
Spine (exo fit)	-0.35	0.05
Shoulder (exo fit)	0.41	0.02

Table 2. Spearman rank correlation to task heart rate differential, ΔHR_{LP}.

	R-Coefficient	P-Value
Gender	-0.43	0.01
Prior Pain (any)	0.33	0.03

The findings do not necessarily indicate a causality, but potential avenues to explore. For example, if broader shoulders led to greater physical exertion, should the shoulders be sized for a more snug fit to reduce exertion or is the correlation due to the stature of the subject, where broader shoulders would lead to greater exertion?

We only had six females from the 33 subjects, which also leads to the need to further explore whether one of the performance criteria of the test method can be gender neutrality.

In the survey responses, subjects indicating shoulder chafing RPDs had an R coefficient of 0.33, with p-value of 0.06, relative to the HR differential, ΔHR_{Final}. Another interesting trend was the subjects who indicated what they least liked about the exoskeleton had an R coefficient of -0.33, with p-value of 0.07. The result is based on converting the free-form response into categories, where the lower values include heat dissipation, leg resistance, and range of motion. The trends indicate more in-depth investigation in how the characteristics of the exoskeleton may impact the physical demands on the subject.

In order to assess the subjects' opinion towards the exoskeleton and the subjects' HR response, a multiway analysis of variance (ANOVA) was computed based on survey responses if the subject agreed that the exoskeleton eased and helped the task or hindered the task. The factors on whether the exoskeleton helped or hindered did not indicate a significant response. The result may indicate the need to reduce the error tolerance in our current response metric.

We intend to further expand our analysis to utilize the video capture and segment the subject to analyze the subject's range of motion based on skeletal joint angles [20]. In cursory visual assessment of the videos, it was noted that subject's often have differing angles in the knee and back when comparing postures between the subject wearing the exoskeleton (squat lifting technique) and the subject without the exoskeleton (stoop lifting technique). The squat lifting technique has been widely regarded as a means to avoid lower back strain, while the stoop technique has cited benefits of ease, reduced energy consumption, and increased stability [21]. We intend to further explore analysis methodologies regarding the support of the exoskeleton for completion of repetitive, physically strenuous tasks, as well as the ability to enable the subject to perform tasks with more ergonomically sound postures.

Conclusions

The NIST Human Subject's study was approved and is underway beginning with the load positioning task test. A Position and Load Test Apparatus for Exoskeletons (PoLoTAE) was designed to include six generic tasks typical of industrial environments, including load handling. Load handling tasks are a current ASTM F48 work item towards a standard test method. The study used a data collection system of cameras, monitors, and HR monitors and the subject completed a survey after the completing the task. Results showed that the combination of HR and other test data from video and the survey provided additional information beyond just HR. For example, subject comfort level may affect HR. Uncertainties were from: data units conversion, subjective exoskeleton fit to the subject by the briefly-trained research team, subject responses to survey questions about exoskeleton fit, HR monitor devices, and subjective review of video and of survey responses. Initial analysis is based on the Spearman rank correlation and results are shown in Tables 1 and 2 with respect to final and task heart rate differential. Trends showed high significance with respect to gender, anthropometry, and fit. As in section VI C, females and those who indicated prior pains generally had lower heart rates in wearing the exoskeleton relative to the baseline. Shoulder width also appeared to have an impact on the subject's physical exertion.. Survey responses of whether the exoskeleton helped or not did not indicate a significant response. However, when asked outside of the survey if the subjects felt the test method exercised the potential capabilities of the exoskeleton, a unanimous affirmative was received by all subjects.

References

[1] DARPA Director's Testimony to the Subcommittee on Terrorism, Unconventional Threats and Capabilities, House Armed Services Committee, United States House of Representatives, March 29, 2006.

[2] Research Exoskeleton Report, http://exoskeletonreport.com/category/research-and-academia/, accessed October 28, 2016.

[3] Brian Lowe, William Billotte, ASTM F48 Formation and Standards for Industrial Exoskeletons and Exosuits, Submission to IISE Transactions on Occupational Ergonomics and Human Factors, Special Issue on Occupational Exoskeletons and Exosuits, Oct 2018.

[4] ASTM F48 Exoskeletons and Exosuits, https://www.astm.org/COMMIT/SUBCOMMIT/F48.htm, accessed Jan 2018.

[5] United States Bureau of Labor, "Employment Projections by Major Industry Sector", Web. 26 Jan. 2017.

[6] A. Caruso, "Statistics of U.S. Businesses Employment and Payroll Summary: 2012", Economy Wide Statistics Briefs, US. Census Bureau, February 2015.

[7] Roger Bostelman and Tsai Hong, Test methods for exoskeletons—lessons learned from industrial and response robotics, Wearable exoskeleton systems: design, control and applications, Chapter 13, pp 335-361, Jan 2, 2018.

[8] NIST Institutional Review Board Study, "Analysis of Exoskeleton-Use for Enhancing Human Performance to Complete Industrial Tasks", June 4, 2018 approval.

[9] Roger Bostelman, Position and Load Test Apparatus for Exoskeletons (PoLoTAE) Design, draft NIST Internal Report, November 2018.

[10] Saad Alabdulkarima, Maury A. Nussbaumb, Influences of different exoskeleton designs and tool mass on physical demands and

Bostelman, Roger; Li-Baboud, YaShian; Virts, Ann; Yoon, Soocheol; Shah, Mili.
"Towards Standard Exoskeleton Test Methods for Load Handling."
Paper presented at WearRAcon 19, Scottsdale, AZ, United States. March 26, 2019 - March 28, 2019.

performance in a simulated overhead drilling task, Applied Ergonomics 74 (2019) 55–66.
[11] NIST Institutional Review Board training video, https://tube.nist.gov/media/t/0_eko39a81, accessed November 11, 2018.
[12] Robert A. Rogers and Roberto Landwehr, The surprising history of the “HRmax = 220 – age” equation, Journal of Exercise Physiology, Vol 5, No. 2, May 2002.
[13] American Heart Association, www.AMA.org, accessed Nov 11, 2018.
[14] Wang, R., Blackburn, G., Desai, M., Phelan, D., Gillinov, L., Houghtaling, P. and Gillinov, M., 2017. Accuracy of wrist-worn heart rate monitors. Jama cardiology, 2(1), pp.104-106.
[15] Parak, Jakub, et al. "Evaluation of the beat-to-beat detection accuracy of PulseOn wearable optical heart rate monitor." Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. IEEE, 2015.
[16] Stahl, Sarah E., et al. "How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough?" BMJ open sport & exercise medicine 2.1 (2016): e000106.
[17] Karvonen, M.J., 1957. The effects of training on heart rate: a longitudinal study. *Ann Med Exp Biol Fenn*, *35*, pp.307-315.
[18] Asselin, P., Knezevic, S., Kornfeld, S., Cirnigliaro, C., Agranova-Breyter, I., Bauman, W.A. and Spungen, A.M., 2015. Heart rate and oxygen demand of powered exoskeleton-assisted walking in persons with paraplegia. *Journal of Rehab. Res. & Dev.*, 52(2), pp.147-160.
[19] Gautheir, T.D., 2001. Detecting trends using spearman's rank correlation coefficient. Environmental forensics, 2(4), pp.359-362.
[20] Omran, M., Lassner, C., Pons-Moll, G., Gehler, P. V., & Schiele, B. (2018). Neural Body Fitting: Unifying Deep Learning and Model-Based Human Pose and Shape Estimation. https://doi.org/10.1109/3DV.2018.00062
[21] Bazrgari B, Shirazi-Adl A, Arjmand N. Analysis of squat and stoop dynamic liftings: muscle forces and internal spinal loads. Eur Spine J. 2006;16(5):687-99.

A-UGV Capabilities

Recommended Guide to Autonomy Levels

Roger Bostelman, Elena Messina
Intelligent Systems Division, Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD, 20899-8230, USA
roger.bostelman, elena.messina@nist.gov

***Abstract*— Automatic-through Autonomous Unmanned Ground Vehicle (A—UGV) has been defined by ASTM Committee F45 as an "Automatic, Automated, or Autonomous vehicle that operates while in contact with the ground without a human operator". However, what do the three "A" levels actually mean to manufacturers, users, or especially potential users? This paper defines, and in many cases provides examples of, recommended autonomy levels for all three automatic, automated, and autonomous unmanned ground vehicles.**

Keywords- A-UGV, autonomous, capability, ASTM F45, classifiers

I. INTRODUCTION

A-UGV (A-unmanned ground vehicle) has been defined by ASTM Committee F45 [1] as an "Automatic, Automated, or Autonomous vehicle that operates while in contact with the ground without a human operator". However, what do the autonomy or capability levels actually mean to manufacturers, users, or especially potential users? Aside from cost, their focus is most likely on A-UGV capabilities, configuration, facility integration, industrial application, and/or many other vehicle functions that will help the user. Potential A-UGV users who search for A-UGVs mainly hear of two current types of systems: automatic guided vehicles (AGVs) - which are preprogrammed vehicles - and mobile robots. "Mobile robot" is an informal term for a vehicle that includes all intelligent functionality beyond automatic guided vehicles. Meanwhile, the term AGV has been expanded over the years to include laser, self, and other guided vehicles. [2] F45 defined the A-UGV term to minimize confusion, first within the committee and second to express a less ambiguous set of terms for the industry. The A-UGV term therefore includes both AGVs and mobile robots, although further definition is required to again limit confusion and misinterpretation since it spans a broad range of capabilities. During the development of the A-UGV term, the following were defined by the F45 terminology task group, although were not formalized as standard:

- automatic-UGV, n—vehicle capable of following a pre-programmed path and that does not deviate from the path without human intervention,
- automated-UGV, n—automatic vehicle with limited ability to deviate from the pre-programmed path,
- autonomous-UGV, n—self-guided vehicle that is able to travel without a pre-programmed path and operates independently to navigate around fixed and moving obstructions.

With these definitions, AGV capabilities, e.g., offboard, pre-planned navigation path segments between waypoints, fit mainly within the 'Automatic' term, whereas mobile robot capabilities can typically fit within all three terms. 'Automated' UGVs can deviate from the originally-planned path, whereas, onboard, continuously-re-planned paths are typical of 'Autonomous' UGVs. Additionally, there are many other functions that can define the Autonomous-UGV's autonomy, such as navigation, docking and software/hardware reconfiguration control based on sensory interaction, knowledge representation, and judgement, and behavior expectations. The combination of the latter functions can also be described as intelligence. As will be described in following sections, the concepts 'autonomy' versus 'intelligence' have been discussed among many groups and for many applications. Briefly, some definitions for autonomy and intelligence are as follows:

Autonomy	Intelligence
Ability to perform intended tasks based on current state and sensing, without human intervention [3]	Ability to acquire and apply knowledge and skills [5]
Self-directing freedom and especially moral independence [4]	1. Ability to learn or understand or to deal with new or trying situations, also the skilled use of reason 2. Ability to apply knowledge to manipulate one's environment or to think abstractly as measured by objective criteria (such as tests) [4]
Freedom from external control or influence; independence [5]	One's capacity for logic, understanding, self-awareness, learning, emotional knowledge, planning, creativity, and problem solving [6]

Other definitions have been developed and provide similar concepts of independence and accomplishing goals based on knowledge and perception of the world. For example, according to [7], to be autonomous, a system must have the capability to independently compose and select among different courses of action to accomplish goals based on its

Bostelman, Roger; Messina, Elena.
"A-UGV Capabilities - Recommended Guide to Autonomy Levels."
Paper presented at 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy. February 25, 2019 - February 27, 2019.

knowledge and understanding of the world, itself, and the situation. However, understanding and perceiving the world and situations can broadly vary.

Unfortunately, the current F45 terminology covers only three levels of autonomy, again suggesting all functions beyond automated fall within autonomous. It is clear that there are further autonomy-level divisions that are needed. As more autonomous industrial vehicles are manufactured and marketed, standard test methods and practices are needed to help inform the user of the expected performance for these advanced capabilities. Consensus-based standards also provide an unambiguous and precise language with which users can specify their required levels of autonomy prior to procuring systems. Laying the groundwork for test methods, a Standard Guide to A-UGV Autonomy-Level would aid the A-UGV user to first understand the variety of vehicle types and capabilities available and to match the advanced vehicle to the advanced task. As opposed to automatic and automated systems, the increased complexity in capability and function also provides increased difficulty in understanding which A-UGV to apply to tasks. To define A-UGV autonomy levels for clear understanding and use by the industrial vehicle industry begins with generically establishing the variety of levels and then fitting their control, capabilities, and functionalities into clear categories.

This paper includes initial sections briefly describing prior efforts towards defining autonomy levels with extended description of Autonomy Levels for Unmanned Systems [8] and then describing the relationship between autonomy and intelligence. This background is essential to allow a focused classification of autonomy for industrial vehicles. This is followed by a section with recommended autonomy levels which includes a table of example industrial vehicle implementation scenarios of each autonomy characteristic for each autonomy level.

II. Prior Efforts Towards Defining Autonomy Levels

There have been numerous prior efforts in defining autonomy levels. One of the most well-known is the ALFUS (Autonomy Levels for Unmanned Systems) effort that has been absorbed within the SAE AS4-D [9]. ALFUS was originally developed by a government informal working group addressing the lack of autonomy measures to support new major Department of Defense programs. The ALFUS development team mined and built upon several other relevant frameworks [10], including: National Aeronautics and Space Administration (NASA) Spacecraft Mission Assessment and Re -planning Tool (SMART) [10], Observe, Orient, Decide, Act (OODA) [11], NIST 4D/Real-time Control System Reference Architecture [27],"Sheridan" Model [12], Defense Science Board Summer Study on Autonomy [13], and others. Some of these, such as the Sheridan model, focused on categorizing the levels of dependence/independence of the automated system from the human.

The resulting ALFUS framework [14] is based on a hierarchical, multi-dimensional model of the main factors that affect autonomy. The three main dimensions (or axes) are: human independence, mission complexity, and environmental complexity. Therefore, the degree of autonomy of a system is characterized not only by how much it relies (or doesn't) on human direction and interaction, but also on the types of tasks it is capable of performing and the types of environments within which it performs them. Each of these axes themselves represent a number of characterization aspects. Detailed discussions and guidance are found in [14], but some examples are:

Mission Complexity potential metrics

- Mission time constraint
- Precision constraints in navigation, manipulation, etc.
- Rules of engagement
- Knowledge requirements in order to plan mission and adjust/adapt to respond to changing conditions

Environmental Complexity potential metrics

- Traversability of terrain (flat clear support surface/floor versus highly uneven, non-uniform)

Visibility

- Dynamicism of environment (moving objects versus static known surroundings)

Human Independence potential metrics

- Scope and range of mission that the system can plan and execute independently
- Ability to generate high-level complex plans versus just derive lower-level plans or signals to system actuators from higher-level plans that were given to it.
- Ability to communicate the relevant information to the appropriate human (including distinguishing between human roles, such as operators, bystanders, adversaries, etc.)

As can be deduced from the examples, the autonomy level for a system is always dependent on the context within which the system performs. Therefore, ALFUS evolved to define a "Contextual Autonomous Capability (CAC) Model for Unmanned Systems." [15][16]

Several other standards that discuss safety and performance of autonomous systems are or have been developed within several standards development organizations. For example, ISO TC 299 Robotics defines autonomy and employs autonomy in several robot standards within several working groups. The International Electrotechnical Commission (IEC) TR 60601-4-1 [18] provides guidance and interpretation of medical electrical equipment and medical electrical systems employing a degree of autonomy. The Institute for Electronic and Electrical Equipment [19] has an active project to develop standard IEEE 7009 for fail-safe design of autonomous and semi-autonomous systems. Given the growing interest in self-driving vehicles, the U. S. Department of Transportation has begun efforts at classifying what they term as automation levels. [17] The aforementioned report references on-road autonomous vehicle taxonomy and definitions in SAE International standard J3016_201609 [20], including three main driving factors: the human driver, the driving automation system, and other vehicle systems. Industrial Truck Standards Development Foundation [21] B56.5 covers safety of automatic guided industrial vehicles and American National Standards Institute (ANSI)/Robotic Industries Association

Bostelman, Roger; Messina, Elena.
"A-UGV Capabilities - Recommended Guide to Autonomy Levels."
Paper presented at 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy. February 25, 2019 - February 27, 2019.

(RIA) [22] 15.08 is developing a mobile robot and mobile manipulator (i.e., robot arm(s) onboard a mobile robot base) safety standard. None of these standards efforts currently provide guidance on the expected operation of industrial autonomous vehicles as is considered in this document and is expected to fall within ASTM F45.

III. The Relationship Between Autonomy and Intelligence

To put autonomy and intelligence into a fairly general example, consider this scenario: babies are relatively intelligent, as compared to adults, born with basic abilities such as reflexes, general motor skills, eating, face matching, and discerning details in the world through their senses. As babies grow, they become more independent or autonomous from care-givers while learning about their environment through experience and education, improving on motor skills, and becoming able to generalize from experiences in order to respond to new situations.

Sometimes, the terms "autonomy" and "intelligence" are used interchangeably. We examine both terms and their relationship as applied to machines, building on the ALFUS high-level summary above. Within the ALFUS framework, the definition of fully autonomous is "a mode of unmanned system (UMS) operation wherein the UMS accomplishes its assigned mission, within a defined scope, without human intervention while adapting to operational and environmental conditions" [23]. Within the scope of ALFUS, intelligence in an unmanned system is defined as its possession of and the ability to exercise contextual autonomous capability [ibid.]. Sanz et al. defined autonomy as the ability of a system to fulfill a task within a given context without external help [24].

A similar perspective is present in the Albus [25] definition of intelligence, initially posed as "the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success, and success is the achievement of behavioral subgoals that support the system's ultimate goal" [26]. In later works which were more application-focused (e.g., [27][28]), the definition was expanded:

- An intelligent system is a system with the ability to act appropriately in an uncertain environment.
- An appropriate action is that which maximizes the probability of successfully achieving the mission goals.
- A mission goal is a desired result that a mission is designed to achieve or maintain.
- A result is represented as a state or some integral measure of a state-time history.
- A mission is the highest-level task assigned to the system.

The Albus definitions of intelligence do not explicitly mention the role of the human in a system's operation. The attribute of being able to independently achieve success is the explicit expression of the autonomy concept. Combining the Albus definition with the Sanz concept "without external help" merges the intelligence and autonomy attributes: "the ability of a system to independently act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success, and success is the achievement of behavioral subgoals that support the system's ultimate goal."

IV. Recommended Autonomy Levels for A-UGVs

For industrial vehicles, A-UGVs have large amounts of human-machine interaction in lower autonomy levels, building to more autonomous functionality having small amounts or no human-machine interaction with increasing A-UGV autonomy. Matching the A-UGV autonomy level to the task may be challenging to the user and therefore, some guidance is warranted. The following sections first define classifiers and recommended autonomy levels for autonomous-UGVs, and then show an example A-UGV classification. Context should also be added to the recommended levels and is then briefly described.

A. Classifiers

Classifiers are a set of terms and their definitions, as shown in Table I, that the A-UGV is capable of performing (e.g., Navigation, Docking, etc.) and that affect the A-UGV performance (e.g., Environmental difficulty, Situation awareness, etc.). Bolded classifiers are defined terms within ASTM F3200-17 [1]. Table 1 defines twelve classifiers specifically focused on A-UGV implementation where the definitions, including those shown in F3200-17, may be different from ones researched in dictionaries. For example, situation awareness is defined in the table by [5]. However, decision-making was modified from existing definitions to be more focused on A-UGV implementation.

TABLE I. AUTONOMY CLASSIFIERS

Classification Category/Metric	**Definition**
Navigation	deciding on and controlling the direction of travel derived from localization and the environment map; see simultaneous localization and mapping (SLAM), localization. DISCUSSION—Navigation can include path planning for location-to-location travel and complete area coverage
Docking	arrival and act of stopping at a position relative to another object
Subtasks	a portion or portions of **tasks** (sequence of movements and measurements that comprise one repetition within a test)
Organization structure	**control**, communications, interaction requirements between the A-UGV and an offboard controller (e.g., central) and/or with other A-UGVs to accomplish desired goal(s)
Decision-making	the action or process of making decisions including the associated system to make the decision (e.g., central control, another A-UGV)
Situation awareness	the perception of environmental elements and events with respect to time or space, the comprehension of their meaning, and the projection of their status after some variable has changed, such as time, or some other variable, such as a predetermined event. [5]
Knowledge requirements	the amount of information and experience necessary to achieve a goal(s)

Bostelman, Roger; Messina, Elena.
"A-UGV Capabilities - Recommended Guide to Autonomy Levels."
Paper presented at 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy. February 25, 2019 - February 27, 2019.

Environmental Difficulty	the A-UGV situation to overcome, deal with, or understand due to natural (e.g., weather, climate, terrain, vegetation), modified (e.g., specific induced environments) and/or observed conditions by the A-UGV during operation
Terrain variation	surface conditions (e.g., ramps, roughness, softness/hardness, etc.) that the A-UGV can traverse
Communication dependencies	reliance upon communication with external A-UGV sources for the A-UGV to achieve goal(s)
Tactical behavior	required A-UGV actions towards a goal(s) beyond the current situation
Human-machine interaction (HMI)	information and action exchanges between human and A-UGV to perform a task by means of a user interface

B. Autonomy Level Guide

A recommended guide, shown in Table II, has been developed that adopts aspects of the referenced autonomy level structures from the "Prior Efforts Towards Defining Autonomy Levels" section and applies a more focused industrial A-UGV perspective that exemplifies expected vehicle performance at each level. Autonomy levels are defined using the classifiers shown along the vertical axis of Table II where the highest level may be, perhaps, a top-level goal for Autonomous – UGVs.

Each level includes a generic definition of that level or groups of levels followed by example capabilities that may fit within that A-UGV level. The first two levels (1 and 2) are defined prior to Table II to allow the A-UGV, at all levels, to be fully or partially controlled by the human operator. The third level (3) is more closely related to the typical automatic guided vehicle (AGV) systems while the fourth level (4) expands the AGV abilities to allow for obstacle detection and avoidance while controlled from the central controller. The term guidepath (and all other bolded terms) is defined in ASTM F3200-17 as the "intended path for an A-UGV used with automatic or automated guidance". The fifth through eighth levels (5 through 8) define autonomous-UGVs. Levels three through eight functionalities are best described in Table II where the table expands autonomous-UGVs across four additional levels to include the minimal (e.g., level five) through maximum (level eight) functionalities. All levels build on previous levels and some level 4 Automatic classifiers simply carry the same functionality from one level to the next with no additional functionality. This is because the Automated-UGV expands only the navigation and docking from the Automatic-UGV.

1. A-UGV (no autonomy)
- Definition: An A-UGV that is controlled only by an A-UGV operator and lacks any autonomy
- Example Capabilities:
  Manual mode, manual control, manual operation of an A-UGV
    - Using an operator control unit to move the A-UGV when not being used in production

2. A-UGV (shared control)
- Definition: Shared control between the A-UGV operator and the A-UGV
- Example Capabilities:
    - A-UGV operator uses human-machine interaction to control minimal A-UGV functionality (e.g., speed) while the A-UGV moves using automated functionality.
    - A-UGV operator is aware of and the A-UGV is not aware of the environment.

3. Automatic-UGV
- Definition: A computer-controlled, unmanned A-UGV that can navigate guidepaths with directed movement by a combination of software and sensor-based guidance systems [30].
- Example Capabilities: (see Table II)
    - Automatic Guided Vehicle (AGV)
    - Guidance for navigation is typically achieved using: laser; embedded wire or magnets in floors; chemical, tape, or other floor markings

4. Automated-UGV
- Definition: A level 3 A-UGV that can also re-plan and navigate away from and return to a guidepath.
- Example Capabilities: (see Table II)

5. through 8. Autonomous-UGV
- Definition: A level 4 A-UGV that can re-plan and navigate without the need for a guidepath and using natural features in the environment.
- Example Capabilities: (see Table II).

V. CONTEXT

Context means the "circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood and assessed". [4] In this case, the definition could be modified as "form the setting for an A-UGV event(s)" and "fully understood and acted upon". [4] In addition to A-UGV autonomy levels, context is also important and includes, for example, the location (e.g., indoor or outdoor) where the vehicle is being used, the task complexity and computation speed required to accomplish the task, the environmental conditions (i.e., bright sun/dark, high heat/extreme cold), and many other possible criteria that place the vehicle in an infinite number of potential situations. Additionally, as the A-UGV moves through its environment, whether factory, hospital, outdoors, or other, the dynamically changing unknowns create an even more complex setting for the A-UGV to complete its task.

Bostelman, Roger; Messina, Elena.
"A-UGV Capabilities - Recommended Guide to Autonomy Levels."
Paper presented at 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy. February 25, 2019 - February 27, 2019.

TABLE II. EXAMPLES OF A-UGV PERFORMANCE FOR RECOMMENDED A-UGV AUTONOMY LEVELS 3 THROUGH 8

	A-UGV CAPABILITY LEVELS					
METRIC	**3 - Automatic**	**4 - Automated**	**5 - Autonomous**	**6 - Autonomous**	**7 - Autonomous**	**8 - Autonomous**
Navigation	levels 1, 2 + follows preprogrammed path	level 3 + leaves path and returns to path, e.g., to avoid an obstacle/A-UGV and returns to path	level 4 + mapping using natural features; finds and self-routes to mainly follow guidepaths, e.g., can deviate from path, not follow initial path	level 5 + no guidepath required, self-routes to waypoints toward goal in intended path areas	level 6 + self-routes to goal along intended or alternative and allowable paths	level 7 + no waypoints required, self-routes to goal along paths using decision-making and value judgement
Docking	levels 1, 2 + stops at preprogrammed waypoints with preprogrammed tolerance	Level 3 + able to dock while off preprogrammed path	level 4 + servo to docking pose in heading, translation, and azimuth	level 5 + servo to docking pose in 6 DoF	level 6 + automatic tolerance variation based on situation, e.g, ± 5 mm dock station alignment, then ± 0.5 mm fine tolerancing with no fixturing	level 7 + dynamic docking with moving objects, e.g., moving agile assembly line (independent vehicle) with 6 DoF low tolerance docking
Subtasks	levels 1, 2 + preprogrammed stop points; e.g., pickup/drop-off loads; pull trailor	same as level 3	level 4 + minimal preplanned stop sequence, e.g., follow lines, edges, paths, and/or no lines or natural features to support navigation through minimally complex areas	level 5 + no preplanned stop sequence, e.g., replan paths and navigation through a complex facility having an unstructured environment with periodically blocked and open paths	level 6 + e.g., follow complex contours of spatially-independent object surfaces using sensory intelligence, adapt speed and position to dynamically acquire suspended loads from overhead cranes and AUVs	level 7 + full decision-making e.g., detect, understand humans vs. objects, and vary speed and functionality based on human vs. object recognition; adjust the payload pose according to delicate handling, vehicle speed, ramps, emergency-stop conditions
Organization structure	levels 1, 2 + central controlled; no collaboration	same as level 3	level 4 + central fleet control + self-controlled; fleet replanning when A-UGV(s) is busy	level 5 + A-UGV-to-AUGV map info. (e.g., obstacle, busy, routes knowledge); on-the-fly route(s)-changes from host take effect during operation	level 6 + self-controlled; send A-UGV commands/route-changes to/from other A-UGVs, e.g., idle A-UGV sent for pickup;	level 7 + integration with other vehicle types (UAV's) and facility equipment (cranes, machine tools)
Decision-making	levels 1, 2 + centrally, offboard -controlled decisions (e.g., A-UGV intersection and zone minding); no self-decisions	same as level 3	level 4 + minimal self-decisions (e.g., pass another A-UGV)	level 5 + moderately complex self-decisions from learned events	level 6 + complex self-decisions from learned events	level 7 + full, self, efficient, real-time planning and execution, highest precision and success rate, maximizes/minimizes on values/cost, benefit/risk.
Situation awareness	levels 1, 2 + none, zone or segment known by central controller	level 3 + detect off-path; continuous plan back to path	level 4 + preplanned route and natural-feature mapping/learning	level 5 + self-planned route; high/low-level learning, e.g., single vs. clustered obstacles, humans vs. obstacles	level 6 + multi-level obstacle grid populated with orders of magnitude obstacles; humans vs. obstacle	level 7 + no obstacle grid required; learn complex situations
Knowledge requirements	levels 1, 2 + digital/analog input/output	same as level 3	level 4 + minimal knowledge/information, e.g., detected obstacles placed in map, infrastructure vs. transients	level 5 + medium knowledge/information, e.g., obstacle prediction from motion	level 6 + obstacle recognition, e.g., humans	level 7 + maximum knowledge/information, e.g., detailed learning (textures, transparency/opaque, soft/hard)
Environmental Difficulty	levels 1, 2 + lights on/off; cold/hot; transitions; e.g., slow/stop when 2D safety sensed obstacles	same as level 3	level 4 + dense obstacle field, e.g., obstacle avoidance	level 5 + navigation, environment sensing unaffected by e.g., sunlight-to-dark, hot-to-cold transition, moderatetly-high humidity, moderate air particle density	level 6 + vary avoidance dependent upon recognized obstacle	level 7 + adaptable to extreme terrain and climate variations and obstacle density and frequency. e.g., high humidity; high air particle density
Terrain variation	levels 1, 2 + moderate friction; flat, hard surface; fine particulates	same as level 3	level 4 + moderately flat surface	level 5 + shallow inclines at any angle	level 6 + to 15% grade, moderate friction ground surface; course particulates	level 7 + outdoor ground surfaces (e.g., soft, rough (> 10 mm dia. stone rubble)); low friction
Communication dependencies	levels 1, 2 + receives commands/monitors progress wireless through central computer; finishes segment upon comm. failure with central source and stops	same as level 3	level 4 + initial comm. reliance with commanded route(s) from central host; no A-UGV control effects with comm. failure although monitor interupts	level 5 + comm. via verbal human commands; e.g., "start route 100", "stop", "pause"	level 6 + complex verbal or gestured human commands; e.g., "take-over A-UGV 4 routes", hand wave, point, etc. to command A-UGV routes; monitor from any wireless comm.	level 7 + fully independent from comm. link; e.g., A-UGV gets initial goals, routes through previous level means with no need for human or host comm. for successful goal
Tactical behavior	none	none	low complexity	level 5 + middle complexity, multi-functional tasks;	level 6 + collaborative, high complexity, multi-functional tasks;	level 7 + highest complexity for all tasks, total independence
Human-machine interaction	levels 1, 2 + maximum HMI	same as level 3	level 4 + HMI for periodic path/task correction	level 5 + infrequent HMI, e.g., stuck with poor plan solution (difficult route)	level 6 + rare HMI, e.g., stuck in extremely complex situation; alert with no self-plan solution	level 7 + no HMI required

Bostelman, Roger; Messina, Elena.
"A-UGV Capabilities - Recommended Guide to Autonomy Levels."
Paper presented at 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy. February 25, 2019 - February 27, 2019.

As in [26], "context impacts the appropriateness of virtually all aspects of an agent's behavior" and "context-sensitivity is fundamental to intelligent behavior". An A-UGV with Level 7 Navigation and Level 6 Subtasks can self-route while replanning paths within a complex facility and an unstructured environment. Context can further provide autonomy level implementation challenges where appropriate tests must be considered to measure the A-UGVs performance. For example, an A-UGV with relatively challenging environmental conditions of, for example, frozen factory walls and floors, bright lights reflecting off the walls and floors, and potentially slick spots on the floor would be completely different than a warm, office-lit, non-slippery floor condition when testing the same autonomy levels. Therefore, the recommendation is that once the autonomy level for a subset or all of the 12 classifiers is determined, the context must also be established and recorded to allow comparison of A-UGV performance to the task. ASTM F3218-17 [1] provides a practice for recording the environmental conditions that can help with recording this aspect of context where the A-UGV is to operate. However, in addition, all other context criteria should also be recorded to capture any environmental effects that might affect the A-UGV performance. Using the previous example, bright lights would be recorded in F3218-17, although bright lights reflecting off shiny, frozen walls and floors may not be recorded on the current form and yet may affect performance of the A-UGV.

VI. ASTM Committee F45 Early Recommendations

A workshop was held as part of the July 2018 ASTM Committee F45 meeting, called: A-UGV Capability Levels. During the workshop, the contents of this paper were presented and discussed. The committee accepted the concept of developing a standard guide to A-UGV capabilities, as opposed to autonomy levels, and is considering an alternative to the examples shown in Table II, beginning with navigation and docking classifiers. The alternatives, shown in Table III, will be discussed in future ASTM F45 meetings.

TABLE III. ASTM COMMITTEE F45 EARLY RECOMMENDATIONS FOR A-UGV CAPABILITY LEVELS FOR NAVIGATION AND DOCKING CLASSIFIERS

	A-UGV CAPABILITY			
CLASSIFIER	**3**	**4**	**5**	**6**
Navigation	Follows preprogrammed path	Leaves preprogrammed path and returns to preprogrammed path	Can find an alternate preprogrammed path	Self-routes to the goal
- infrastructure dependence	relies on infrastructure	relies on infrastructure	relies on infrastructure	relies on infrastructure
	does not rely on infrastructure; corrects for errors	does not rely on infrastructure; corrects for errors	does not rely on infrastructure; corrects for errors	does not rely on infrastructure; corrects for errors
Docking	Docks at preprogrammed waypoints	Able to adjust based on local docking position	Dynamic docking with moving objects	
- infrastructure dependence	relies on infrastructure	relies on infrastructure	relies on infrastructure	
	does not rely on infrastructure; corrects for errors	does not rely on infrastructure; corrects for errors	does not rely on infrastructure; corrects for errors	
- docking degrees of freedom	x (heading)	x (heading)	x (heading)	
	y (side-to-side)	y (side-to-side)	y (side-to-side)	
	z (vertical)	z (vertical)	z (vertical)	
	roll (rot. about x)	roll (rot. about x)	roll (rot. about x)	
	pitch (rot. about y)	pitch (rot. about y)	pitch (rot. about y)	
	yaw (rot. about z)	yaw (rot. about z)	yaw (rot. about z)	

VII. Conclusions

This document is intended to provide a broad overview of the nature of defining and categorizing autonomy for industrial vehicles. It builds upon existing work that explored the many dimensions of autonomy for unmanned systems. This effort focuses on the A-UGV domain specifically, seeking to clarify the nomenclature of Automatic, Automated, or Autonomous vehicle. Users can use this framework to use a more functionality-based method with autonomy metrics to describe the advanced capabilities of autonomous-UGVs beyond a single autonomous-UGV category. Manufacturers of A-UGVs can more fully describe their vehicle's capabilities by identifying the autonomy levels for various capabilities, as well as for the overall A-UGV. This framework can provide standards committees, such as ASTM F45, and possibly ITSDF B56.5, and RIA 15.08, a means to expand their standards development roadmap, incorporating relevant aspects of the autonomy categories.

References

[1] ASTM Committee F45 Driverless Automatic Guided Industrial Vehicles, www.astm.org/F45, accessed January 26, 2018.

[2] Josh Bond, "AGVs: Predictably Flexible", http://www.mmh.com/article/agvs_predictably_flexible, January 19, 2018, Accessed January 26, 2018.

[3] International Organization of Standards (ISO) 8373:2012 Robots and robotic devices – Vocabulary, 2012.

[4] Merriam-Webster, https://www.merriam-webster.com/dictionary/, accessed 1/24/2018.

[5] Google, www.google.com, accessed 1/24/2018.

[6] Wikipedia, www.wiki.com, accessed 1/24/2018.

[7] L.G. Shattuck, "Transitioning to Autonomy: A human systems integration perspective", Presentation at Transitioning to Autonomy: Changes in the role of humans in air transportation, March 11, 2015. Available at human-actors.arc.nasa.gov/workshop/autonomy/download/presentations/Shaddock%20.pdf (Accessed November 3, 2017.)

[8] Hui-Min Huang, Kerry Pavek, James Albus, Elena Messina, "Autonomy Levels for Unmanned Systems (ALFUS) Framework: An Update", 2005 SPIE Defense and Security Symposium, Orlando, Florida, 2005.

[9] SAE AS4-D Unmanned Systems Steering Committee, https://www.sae.org/servlets/works/committeeHome.do?comtID=TEAAS4, accessed November 2, 2017.

[10] Ryan W. Proud, Jeremy J. Hart, and Richard B. Mrozinski, "Methods for Determining the Level of Autonomy to Design into a Human Spaceflight Vehicle: A Function Specific Approach, https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100017272.pdf, accessed November 2, 2017.

[11] Lieutenant Colonel Jeffrey N. Rule, "A Symbiotic Relationship: The OODA Loop, Intuition, and Strategic Thought", Strategy Research Project, US Army War College, http://www.dtic.mil/dtic/tr/fulltext/u2/a590672.pdf, accessed January 26, 2018.

[12] Sheridan, T.B., and Verplank, W. L., Human and Computer Control for Undersea Teleoperators. 1978, MIT Man-Machine Systems Laboratory.

[13] Defense Science Board Summer Study, https://www.hsdl.org/?view&did=794641, June 2016, accessed November 2, 2017.

[14] Hui-Min Huang and Elena Messina, "Autonomy Levels for Unmanned Systems (ALFUS) Framework Volume II: Framework Models Initial Version," National Institute of Standards and Technology Special Publication 1011-II-1.0, December 2007.

[15] Hui-Min Huang, Kerry Pavek, Mark Ragon, Jeffry Jones, Elena Messina, James Albus, "Characterizing Unmanned System Autonomy:

Bostelman, Roger; Messina, Elena.
"A-UGV Capabilities - Recommended Guide to Autonomy Levels."
Paper presented at 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy. February 25, 2019 - February 27, 2019.

Contextual Autonomous Capability and Level of Autonomy Analysis," Proceedings of the 2007 SPIE Defense and Security Symposium Unmanned Systems Technology IX, Orlando, FL, April 2007.

[16] Hui-Min Huang, H., Elena Messina, Adam Jacoff, Robert Wade, Michael McNail, "Performance Measures Framework for Unmanned Systems (PerMFUS): Models for Contextual Metrics," Proceedings of the 2010 Performance Metrics for Intelligent Systems Workshop (PerMIS'10), NIST Special Publication 1113, September 2010.

[17] "Automated Driving Systems 2.0: A Vision for Safety", DOT HS 812 442, Version 9a, 13069a, September 6, 2017.

[18] IEC TR 60601-4-1:2017 Medical electrical equipment -- Part 4-1: Guidance and interpretation -- Medical electrical equipment and medical electrical systems employing a degree of autonomy, 2017.

[19] Fail-Safe Design of Autonomous and Semi-Autonomous Systems, https://standards.ieee.org/develop/project/7009.html, accessed 3 February 2018.

[20] Standard J3016_201609 Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, https://www.sae.org/standards/content/j3016_201609/, accessed 3 February 2018.

[21] Industrial Truck Standards Development Foundation B56.5 Driverless Automatic Guided Industrial Vehicles, www.itsdf.org, accessed 3 February 2018.

[22] American National Standards Institute/Robotic Industry Association 15.08, Mobile Robot Safety, www.robotics.org, accessed 3 February 2018.

[23] Hui-Min Huang, Ed., "Autonomy Levels for Unmanned Systems, Volume I: Terminology, Version 1.1," NIST Special Publication 1011, September 2004.

[24] R. Sanz, F. Matia and S. Galan, "Fridges, elephants, and the meaning of autonomy and intelligence," Proceedings of the 2000 IEEE International Symposium on Intelligent Control. Held jointly with the 8th IEEE Mediterranean Conference on Control and Automation (Cat. No.00CH37147), Rio Patras, 2000, pp. 217-222. doi: 10.1109/ISIC.2000.882926

[25] James S. Albus, "Outline for a theory of intelligence," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 473-509, May/Jun 1991. doi: 10.1109/21.97471.

[26] Hui-Min Huang, "Autonomy Levels for Unmanned Systems (ALFUS)", Presentation to the ALFUS Working Group, SAE AS4D Committee, 2005.

[27] Albus, J., Huang, H., Messina, E., Murphy, K., et al., "4D/RCS Version 2.0: A Reference Model Architecture for Unmanned Vehicle Systems," NIST Interagency/Internal Report (NISTIR) 6910, August, 2002.

[28] Meystel, A.M. and E.R. Messina. Measuring the Performance and Intelligence of Systems. Proceedings of Performance Measurement of Intelligent Systems (PerMIS), Gaithersburg, MD, 2000.

[29] Turner, RM. Context-sensitive reasoning for autonomous agents and cooperative distributed problem solving. In Proceedings of the IJCAI Workshop on Using Knowledge in its Context 1993 (pp. 141-151).

[30] Material Handling Industry of America, http://www.mhi.org/fundamentals/automatic-guided-vehicles, accessed November 2, 2017.

Bostelman, Roger; Messina, Elena.
"A-UGV Capabilities - Recommended Guide to Autonomy Levels."
Paper presented at 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy. February 25, 2019 - February 27, 2019.

Proceedings of the ASME 2019 14th International Manufacturing Science and Engineering Conference
MSEC 2019
June 10-14, 2019, Erie, PA, USA

MSEC2019-2896

ON THE IMPACT OF WIRELESS COMMUNICATIONS ON CONTROLLING A TWO-DIMENSIONAL GANTRY SYSTEM

Mohamed Kashef and Richard Candell*
National Institute of Standards and Technology,
Gaithersburg, Maryland
Email: {mohamed.kashef, richard.candell}@nist.gov

Sebti Foufou
University of Burgundy,
Dijon, France
Email: sfoufou@u-bourgogne.fr

ABSTRACT

Industrial wireless is essential to achieve the vision of future manufacturing systems which are highly dynamic and reconfigurable, and communicate large amounts of data. Main challenges of wireless deployment include the stochastic nature of the wireless channels and the harsh industrial transmission environment. In this work, a typical two-dimensional gantry system is controlled by a local controller which receives G-code commands wirelessly over a Wi-Fi network. The industrial wireless channel is replicated using a radio frequency (RF) channel emulator where various scenarios are considered and various wireless channel parameters are studied. The movement of the gantry system tool is tracked using a vision tracking system to quantify the impact of the wireless channel on the system performance. Numerical results are presented including the total run time of an industrial process and the dwell times at various positions through the process.

Introduction

In future manufacturing systems, wireless communications technology plays an important role in achieving flexibility and scalability through allowing the communications between larger numbers of sensors and actuators and allowing more flexible mobility of equipment. The use of wireless communications in factory automation faces various challenges including the delay and reliability requirements, and the harsh industrial radio frequency (RF) environments [1]. Most of the established industrial wireless technologies such as WirelessHART and ISA100.11a are developed for low data rate process automation applications [2]. On the other hand, home and office wireless technologies such as Bluetooth and Wi-Fi can be used for some high data rate industrial applications depending on their requirements [3].

Gantry systems are widely adopted in various industrial applications where they can be used to hold and position a variety of tools for different purposes. Examples of applications include electronic boards assembly, material handling, sorting, scanning, pick and place, welding, cutting, and plotting. Typically, the motor controller is connected using wires to the gantry system motors for path control. However, the task commands and supervisory control in a factory work-cell can be initiated from a supervisory controller which manages the work-cell process through collecting inputs from various sensors and equipment. As a result, the use of wireless communications between the supervisory controller and the gantry system controller has the following advantages. First, it allows for more efficient operation by connecting the supervisory controller to various entities wirelessly and hence better control strategies are applied [4]. The wireless connection allows for the use of the gantry system in various applications and locations in reconfigurable work-cells [5].

*Mohamed Kashef and Richard Candell have contributed equally to this work.

Hany, Mohamed; Candell, Richard; Foufou, Sebti.
"ON THE IMPACT OF WIRELESS COMMUNICATIONS ON CONTROLLING A TWO-DIMENSIONAL GANTRY SYSTEM."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

The use of wireless communications in controlling various gantry systems has been considered before in the literature such as in [4], [6–8]. In [6], a centralized supervisory controller is connected to multiple gantry machines over a Wi-Fi wireless network. The goal of [6] is implementing the interface between Wi-Fi and various wired protocols. In [7], the performance of a gantry system is studied where a method for testing the delay time impacts is proposed to compare various delay sources including the wireless transmissions. In [4], the control of gantry cranes for various applications is discussed where the advantages, system architecture, and challenges are considered. In [8], a pick and place robot is controlled through Zigbee wireless transmissions. All the commands are transfered to the robot controller over the wireless network.

In this work, we evaluate the performance of a two dimensional gantry system that receives commands through a Wi-Fi wireless network. The two dimensional gantry system is a commercial one which operates through G-code commands to its controller. The G-code commands are streamed using a computer to let the gantry system perform move and wait actions in a predetermined path. The commands are transmitted over a Wi-Fi, IEEE802.11n, wireless network. All the Wi-Fi transmitters and receivers are connected to an RF channel emulator to replicate the industrial channel characteristics and control various RF channel parameters. The RF channel emulator is capable of varying the wireless link distance, channel path loss exponent, the channel impulse response (CIR), and the shadowing variance. The position of the two-dimensional gantry system is recorded using a high-resolution vision tracking system. The vision system has high resolution in both distance and time and is connected to the G-code generating computer to synchronize the experiment start time.

The remainder of the paper is organized as follows. In the following section, the system model is considered where the testbed hardware and the experiment scenario are explained in detail. Then, the data collection strategy is illustrated and the performance criteria are defined. Later, the experimental results are presented and discussed. Finally, conclusions are drawn.

System Model

In this section, we introduce the experimental setup including all the hardware components and the operating conditions. We also identify the information flow through the testbed.

Testbed Architecture The testbed is composed of four main components related to the operation of the gantry system. These components are the gantry system, the Wi-Fi network, the RF channel emulator, and the supervisory computer. First, the gantry system is a commercial X-carve machine that was made initially for engraving various types of materials [9]. In our testbed, we allowed the tool to move in two directions only, with a velocity of 8000 millimeter (mm) per minute in each direction. The position resolution of the machine is around 0.1 mm. The motors of the gantry systems are controlled through a wired connection with the controller that receives G-code commands [10] through a Universal Serial Bus (USB) port. The gantry system controller has a buffer that allows it to store up to 16 G-code commands for the continuity of task execution.

Second, the supervisory computer is responsible for generating and sending the G-code commands to the gantry system controller. In this testbed, we used a generic G-code sender in order to stream the G-code commands over the wireless network. We used an Intel Bay Trail Next Unit of Computing (NUC) with Intel N2930 Celeron processor, 8 GB Random Access Memory (RAM), and 64-bit Windows 7 operating system. It has a Qualcomm Atheros AR946x wireless network adapter with two attached antennas [11].

Third, a Wi-Fi network used for wireless data transfer is composed of the supervisory computer Wi-Fi interface, a router, and a USB to Wi-Fi converter. We have used a TP-link Wi-Fi, IEEE802.11n, commercial router. The supervisory computer has its own Wi-Fi interface originally. The gantry system controller is connected to the Wi-Fi network through a USB to Wi-Fi interface. It includes two components, namely, 1) Antaira 1-port IEEE802.11b/g/n wireless serial device server [12] which converts RS232 serial data into Wi-Fi signal, and 2) an RS232 to USB conversion which is done through running a script over a computer.

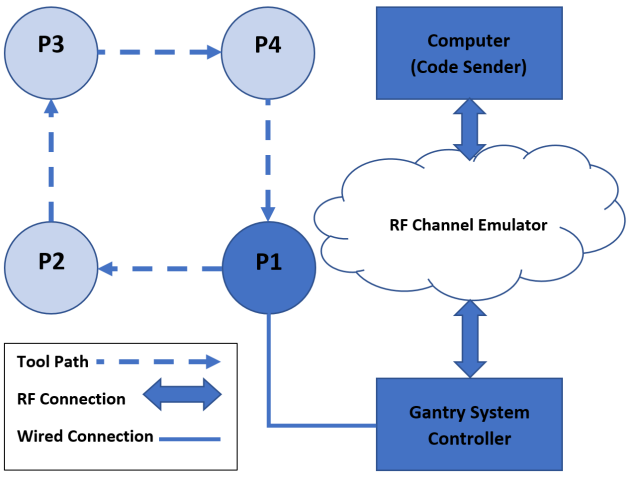


FIGURE 1. Testbed Architecture

Finally, a radio frequency channel emulator is used to repli-

Hany, Mohamed; Candell, Richard; Foufou, Sebti.
"ON THE IMPACT OF WIRELESS COMMUNICATIONS ON CONTROLLING A TWO-DIMENSIONAL GANTRY SYSTEM."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

cate the multi-path and path loss environment for a mesh network of up to 8 physical nodes and 56 virtual links between those nodes. The used channel emulator is RFnest D508 and the corresponding software is RFview [13]. The channel emulator supports an instantaneous bandwidth of 250 MHz (4 nanosecond tap spacing) with an effective dynamic range of 73 dB that includes all analog and digital realization impacts. The emulator is controlled by a nearby computer which loads the path loss model and channel impulse response for each communications link.

Experimental Setup In the experimental study, we run a scenario in which the gantry tool moves sequentially between the four positions P1, P2, P3, and P4 shown in Fig. 1. The horizontal distance between P1 to P2 and P3 to P4 is 8 inches, while the vertical distance between P2 to P3 and P4 to P1 is 10 inches. The gantry system has different dwell times at each of the positions. The dwell times are 2, 1, 0.5, and 0.1 seconds for P1, P2, P3, and P4, respectively.

Each run of the above described scenario continues for 30 complete cycles over all the four positions. This scenario replicates the applications of gantry systems in various fields. It can be similar to a scenario when the gantry tool picks an item at P1 then moves it for processing at the two other positions P2 and P3. Finally, the gantry tool places the item at P4. The pick and place application can be repeated continuously over time, but we make it with a finite cycle to test various wireless channel parameters.

On the other hand, the wireless channel impact is produced through the RF channel emulator. Based on node positions, line of sight (LOS) and non-LOS (NLOS) channels may exist. LOS channels are typically characterized, compared to NLOS channels, by a lower loss exponent and a channel impulse response (CIR) with more energy on the first peak (higher K-factor). First, we consider the benchmark channel with free-space log-distance path loss and ideal CIR which has no multi-path. Second, we consider a measured delay profile of an industrial environment where the CIR is experimentally measured and processed to be deployed using the channel emulator and to reflect the industrial environment impact [1]. While deploying the measured channels, the emulator produces random Rayleigh fading channels for the NLOS cases and random Rician fading channels for the LOS cases following the injected CIRs. Moreover, time-varying log-normal shadowing is introduced due to the fluctuations in the signal level because of obstructions. The variance of zero-mean log-normal shadowing is set through the emulator.

Data Analysis

In this section, we describe the process of collecting the gantry system tool position information. Then, data processing steps are explained in detail where various performance criteria are formulated.

Data Acquisition In order to collect the position information of the gantry system tool, we used a vision tracking system. We deployed the OptiTrack V120 Trio with three cameras, frame rate of 120 frames per second, and 640 x 480 resolution per camera. The tracking system is connected to a computer for data acquisition at which the real-time position information is tracked and stored. We used the supervisory computer for controlling the vision system in order to synchronize the beginning of tracking data acquisition to the gantry system experiment start. The data acquisition setup is shown in Fig. 2.

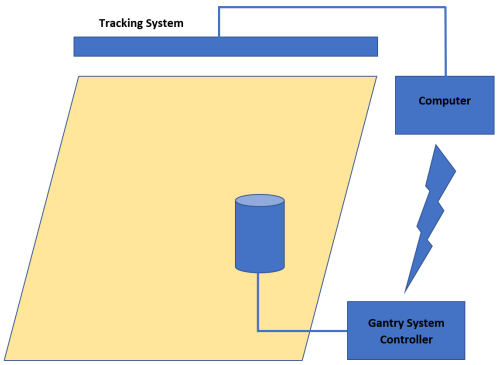


FIGURE 2. Data Acquisition Setup

Data Processing Let's define any tracked position by the pair (x,y) and the preset positions, P_i where $i \in \{1,2,3,4\}$, by (x_i,y_i). The vision tracking system produces a list of position values at a rate of 120 Hz. First, we need to calculate the dwell time at every position P_i during the gantry system movement sequence. We consider the gantry system tool in the position P_i when it is measured to fall in a circle with the center of P_i and a radius of R. The condition is defined as follows

$$\sqrt{(x-x_i)^2+(y-y_i)^2} \leq R. \quad (1)$$

We set the number of runs at each channel setting to M and the number of cycles per run to N. The dwell times sequence at P_i in the mth run is denoted by $d_i(m)$ which is defined as the vector of the times over which the gantry system tool stayed within a distance R of P_i without interruption. The elements of the dwell times vector are denoted by $d_i(m,n)$ where n is the index of the tool being at the position P_i at each cycle. The average dwell time

Hany, Mohamed; Candell, Richard; Foufou, Sebti.
"ON THE IMPACT OF WIRELESS COMMUNICATIONS ON CONTROLLING A TWO-DIMENSIONAL GANTRY SYSTEM."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

for each position, denoted by $\bar{d}_i$, is then calculated as follows

$$\bar{d}_i = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} d_i(m,n). \quad (2)$$

The preset dwell times for various positions are denoted by D_i which equals 2, 1, 0.5, 0.1 seconds, respectively. The average normalized absolute error in the dwell time for P_i, which is denoted by $\tilde{e}_i$, is then expressed as follows

$$\tilde{e}_i = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{|d_i(m,n) - D_i|}{D_i}, \quad (3)$$

and the average normalized absolute error in the dwell time over all positions is calculated as follows

$$\tilde{e} = \frac{\tilde{e}_1 + \tilde{e}_2 + \tilde{e}_3 + \tilde{e}_4}{4}. \quad (4)$$

Another performance criterion is the average total run time of the experiment which is denoted by T. The total run time for the mth run is denoted by $T(m)$. The value of $T(m)$ is the period of time since the run starts with the gantry system tool at P_1 and ends after all the N cycles with the gantry system back to P_1. Then, the value of T is calculated as follows

$$T = \sum_{m=1}^{M} T(m) \quad (5)$$

Finally, we also evaluate the number of incomplete runs, denoted by M_I, which is defined as the number of runs at a specific setting where the run has not reached its end due to the wireless network drop.

Numerical Results

In this section, we present the results obtained by running the experiment for various system and channel parameters. We set the distance between the gantry system and the supervisory computer to {20, 25, 30, 35} meters. The shadowing variance takes the values {0, 20, 40, 60} dB. We use three different wireless channel settings over the RF channel emulator, namely, free-space, LOS, and NLOS. Free-space represents the ideal channel with loss exponent of 2 and ideal CIR with no multipath or fading. The LOS and NLOS channel parameters are set according to the measurements in [1]. In the case of LOS, the loss exponent is 1.8 and the CIR taps vary according to a Ricean distribution. In the case of NLOS, the loss exponent is 2.7 and the CIR taps vary according to a Rayleigh distribution.

In this work, we set the number of runs for each setting to $M = 5$. We set the number of cycles per run to $N = 30$. The total run time of the experiment is measured to be around 347 seconds when no communications errors occur and the dwell times at P_1, P_2, P_3, and P_4 are set to 2, 1, 0.5, 0.1 seconds, respectively. The radius R is set to 0.25 millimeter. It is important to note that the Wi-Fi network may be dropped if certain data are missed or delayed which may randomly happen. Also, the effects of missing various types of data are different and hence may cause longer wait times in some cases than the others. In the following subsections, we show the numerical results of various performance criteria and highlight the values which deviate considerably from the typical values.

Total Run Time The typical run time is about 347 seconds, but due to the wireless channel imperfections, the total run time increases when the gantry system tool stops at certain positions longer than required because no new commands are received by the gantry system controller. First, the performance of LOS and NLOS is almost perfect for short distances and low shadowing variance values. The highlighted values are the ones which deviated from the typical value of the system run time. In most practical scenarios, shadowing cannot be more than 20 dB and hence the performance is acceptable for distances up to al-

TABLE 1. Total Run Time in Seconds

Dist.	Shadowing	Free-space	LOS	NLOS
20 m	0 dB	347.0067	347.2533	346.8383
	20 dB	348.0033	347.0633	356.5867
	40 dB	348.2983	353.6729	403.2833
	60 dB	357.2417	427.5389	-
25 m	0 dB	347.2933	346.9800	347.7875
	20 dB	347.3133	346.9708	349.2208
	40 dB	376.0967	354.5563	436.4000
	60 dB	358.0167	375.2278	-
30 m	0 dB	347.3033	347.2567	346.9229
	20 dB	347.675	347.1583	407.6313
	40 dB	348.4861	373.6833	364.9083
	60 dB	361.1958	375.6292	-
35 m	0 dB	347.6233	347.1167	347.6542
	20 db	347.7083	352.6542	377.6104
	40 dB	424.925	375.3983	467.3500
	60 dB	354.2	370.5375	-

Hany, Mohamed; Candell, Richard; Foufou, Sebti.
"ON THE IMPACT OF WIRELESS COMMUNICATIONS ON CONTROLLING A TWO-DIMENSIONAL GANTRY SYSTEM."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

most 30 meters. All the runs at NLOS with 60 dB shadowing were not completed.

Average Normalized Error The average normalized error in dwell times represents the average of the normalized deviations of dwell times at all four positions of the gantry systems. In the following subsection the average values of the dwell times are shown. Due to the data acquisition scheme, the typical value of the average normalized error is around 0.26. Depending on the type of dropped packets, the network is dropped or the gantry system tool is stopped at a location for a long time. As a result, the errors are generally higher for poor channels, but may take exceptionally high values at certain settings.

TABLE 2. Average Normalized Error

Dist.	Shadowing	Free-space	LOS	NLOS
20 m	0 dB	0.2660	0.2628	0.2767
	20 dB	0.2665	0.2666	0.3951
	40 dB	0.2721	0.3424	0.4279
	60 dB	0.2641	1.2421	-
25 m	0 dB	0.2665	0.2618	0.2784
	20 dB	0.2628	0.2612	0.6432
	40 dB	0.3728	0.6581	0.9981
	60 dB	0.9803	0.7500	-
30 m	0 dB	0.2643	0.2705	0.2744
	20 dB	0.2593	0.2645	1.3256
	40 dB	0.3436	1.0714	0.5307
	60 dB	0.3937	0.8598	-
35 m	0 dB	0.2636	0.2620	0.2779
	20 dB	0.4979	0.2799	2.1956
	40 dB	0.9447	0.4860	0.6589
	60 dB	0.5029	0.4153	-

Average Dwell Times at Various Positions In this subsection, we present the average dwell times at all four positions of the gantry system. In certain cases where the gantry system tool performs a task which requires controlled timing, the increase of dwell time is considered a major problem in these applications. As a result, the wireless network has to operate within settings that guarantee satisfactory performance.

TABLE 3. Average Dwell Time in Seconds at P_1

Dist.	Shadowing	Free-space	LOS	NLOS
20 m	0 dB	2.0825	2.0805	2.0791
	20 dB	2.0811	2.0808	2.0763
	40 dB	2.0765	2.0685	3.1047
	60 dB	2.0751	2.2285	-
25 m	0 dB	2.0803	2.0809	2.0781
	20 dB	2.0803	2.0807	2.3019
	40 dB	2.9579	2.0780	2.2129
	60 dB	1.9876	2.0876	-
30 m	0 dB	2.0805	2.0801	2.0796
	20 dB	2.0793	2.0795	2.9043
	40 dB	2.0772	2.4617	2.1381
	60 dB	2.0755	2.2027	-
35 m	0 dB	2.0799	2.0784	2.0785
	20 dB	1.8011	2.2293	2.1386
	40 dB	3.5654	2.0776	4.5896
	60 dB	2.0758	2.0759	-

TABLE 4. Average Dwell Time in Seconds at P_2

Dist.	Shadowing	Free-space	LOS	NLOS
20 m	0 dB	1.0793	1.0769	1.0749
	20 dB	1.0783	1.0781	1.0743
	40 dB	1.1027	1.0889	1.0745
	60 dB	1.0763	1.0750	-
25 m	0 dB	1.0784	1.0784	1.0764
	20 dB	1.0787	1.0773	1.0750
	40 dB	1.0773	1.0765	3.4386
	60 dB	1.1220	1.3669	-
30 m	0 dB	1.0781	1.0780	1.0759
	20 dB	1.0768	1.0776	1.0754
	40 dB	1.0759	1.0761	2.0691
	60 dB	1.1293	1.0953	-
35 m	0 dB	1.0774	1.0777	1.0750
	20 dB	1.1994	1.0773	1.0720
	40 dB	1.5324	1.9765	1.0709
	60 dB	1.3594	1.0851	-

Hany, Mohamed; Candell, Richard; Foufou, Sebti.
"ON THE IMPACT OF WIRELESS COMMUNICATIONS ON CONTROLLING A TWO-DIMENSIONAL GANTRY SYSTEM."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

TABLE 5. Average Dwell Time in Seconds at P_3

Dist.	Shadowing	Free-space	LOS	NLOS
20 m	0 dB	0.5839	0.5826	0.5838
	20 dB	0.5828	0.5826	0.8248
	40 dB	0.5827	0.7386	0.5832
	60 dB	0.5829	0.5814	-
25 m	0 dB	0.5826	0.5834	0.5846
	20 dB	0.5826	0.5826	1.1497
	40 dB	0.5822	0.7936	0.8091
	60 dB	0.6009	0.5819	-
30 m	0 dB	0.5824	0.5953	0.5851
	20 dB	0.5825	0.5829	1.3481
	40 dB	0.5814	0.5824	0.5823
	60 dB	0.6525	0.5902	-
35 m	0 dB	0.5807	0.5843	0.5842
	20 dB	0.6486	0.5841	0.6347
	40 dB	0.5810	0.5833	0.7065
	60 dB	0.9295	0.8857	-

TABLE 6. Average Dwell Time in Seconds at P_4

Dist.	Shadowing	Free-space	LOS	NLOS
20 m	0 dB	0.1776	0.1769	0.1825
	20 dB	0.1782	0.1783	0.1818
	40 dB	0.1782	0.1769	0.1918
	60 dB	0.1777	0.5616	-
25 m	0 dB	0.1782	0.1761	0.1829
	20 dB	0.1767	0.1762	0.2047
	40 dB	0.1771	0.2930	0.1829
	60 dB	0.4591	0.3426	-
30 m	0 dB	0.1774	0.1774	0.1812
	20 dB	0.1756	0.1775	0.4079
	40 dB	0.2097	0.4814	0.1820
	60 dB	0.2103	0.4062	-
35 m	0 dB	0.1776	0.1763	0.1829
	20 dB	0.2396	0.1759	0.9372
	40 dB	0.3302	0.1762	0.1857
	60 dB	0.1755	0.1767	-

Incomplete Runs Finally, we present the number of incomplete runs out of five for each of the settings. The network drop is not related to the period of communication link interruption, but it is related to the type of packets dropped. However, having higher shadowing variance is a main cause of dropping the network connectivity compared to the channel distance and CIR characteristics.

TABLE 7. The Number of Incomplete Runs at Each Setting

Dist.	Shadowing	Free-space	LOS	NLOS
20 m	0 dB	0	0	0
	20 dB	0	0	0
	40 dB	0	1	2
	60 dB	3	2	5
25 m	0 dB	0	0	1
	20 dB	0	1	3
	40 dB	0	1	0
	60 dB	3	2	5
30 m	0 dB	0	0	1
	20 dB	1	0	1
	40 dB	2	1	2
	60 dB	3	3	5
35 m	0 dB	0	1	1
	20 dB	1	1	1
	40 dB	3	0	4
	60 dB	2	3	5

Conclusions

In this paper, we introduced an experimental study of the impacts of wireless channel parameters on a gantry system performance. We introduced the testbed architecture, the experiment setup, and performance measurements. We deployed a vision tracking system for collecting position data from the gantry system. We have shown that the main wireless impact on the performance is longer dwell times at certain positions of the gantry system tool path. The used gantry system has a G-code commands buffer which reduced the abrupt variations on the performance. The system performance is clearly different for LOS and NLOS settings. Generally, wireless technology benefits can be reaped on similar systems by designing the system appropriately based on the RF environment of the gantry system. Design parameters include buffer size, maximum transmission range, and requirements on LOS existence.

Hany, Mohamed; Candell, Richard; Foufou, Sebti.
"ON THE IMPACT OF WIRELESS COMMUNICATIONS ON CONTROLLING A TWO-DIMENSIONAL GANTRY SYSTEM."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

ACKNOWLEDGMENT

We would like to thank Rushad Antia for his efforts in building the testbed and setting up the supervisory computer.

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

[1] Candell, R., Remley, C., Quimby, J., Novotny, D., Curtin, A., Papazian, P., Koepke, G., Diener, J., and Kashef, M., 2017. Industrial wireless systems: Radio propagation measurements. Tech. rep., National Institute of Standards and Technology, Gaithersburg, MD.

[2] Nixon, M., 2012. A comparison of wirelesshart and isa100.11a. Tech. rep., MSU-CSE-06-2; Emerson Process Management, Round Rock, TX, USA.

[3] Bregulla, M., Feucht, W., Koch, J., de Mur, G., Schade, M., Weczerek, J., and Wenzel, J., 2011. Wireless solutions in automation. ZWEI, March.

[4] Wireless control of gantry cranes in industries. EMNICS Technologies Pvt Ltd, 2015.

[5] Gaspar, T., Ridge, B., Bevec, R., Bem, M., Kova, I., Ude, A., and Gosar, Z. "Rapid hardware and software reconfiguration in a robotic workcell". In 2017 18th International Conference on Advanced Robotics (ICAR), pp. 229–236.

[6] Al-Saedi, I. R. K., Mohammed, F. M., and Obayes, S. S., 2017. "CNC machine based on embedded wireless and internet of things for workshop development". In 2017 International Conference on Control, Automation and Diagnosis (ICCAD), pp. 439–444.

[7] Wozniak, A., and Jankowski, M., 2016. "Wireless communication influence on CNC machine tool probe metrological parameters". *The International Journal of Advanced Manufacturing Technology,* ***82***(1), Jan, pp. 535–542.

[8] Kulkarni, A. D., Sujatha, K., and Shinde, N., 2015. "Article: A pick and place robot: A flexible robot with adjustable rubber hand". *International Journal of Computer Applications,* ***128***(8), October, pp. 16–19. Published by Foundation of Computer Science (FCS), NY, USA.

[9] X-carve [online]. https://www.inventables.com/technologies/x-carve. Accessed: 08/10/2018.

[10] Smid, P., 2008. *CNC Programming Handbook*. 3rd Edition, ISBN 9780831133474.

[11] Industrial Intel Bay Trail Fanless NUC computer ML100G-10. https://www.logicsupply.com/ml100g-10/. Accessed: 2018-08-13.

[12] Antaira stw-611c: Product specifications. http://www.antaira.com/products/STW-611C. Accessed: 2018-08-13.

[13] Rfnest: product specifications. http://www.i-a-i.com/wp-content/uploads/2017/07/RFnest-Specsheet-2017.pdf. Accessed: 2017-08-14.

Hany, Mohamed; Candell, Richard; Foufou, Sebti.
"ON THE IMPACT OF WIRELESS COMMUNICATIONS ON CONTROLLING A TWO-DIMENSIONAL GANTRY SYSTEM."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

Proceedings of the ASME 2019 14th International Manufacturing Science and Engineering Conference
MSEC2019
June 10-14, 2019, Erie, PA, USA

MSEC2019-2902

DEVELOPING MEASUREMENT SCIENCE TO VERIFY AND VALIDATE THE IDENTIFICATION OF ROBOT WORKCELL DEGRADATION

Brian A. Weiss
National Institute of Standards and Technology (NIST)
Gaithersburg, Maryland, USA

ABSTRACT

Robot systems have become more prevalent in manufacturing operations as the technology has become more accessible to a wider range of manufacturers, especially small to medium-sized organizations. Although these robot technologies have become more affordable, easier to integrate, and greater in functional capability, these advanced systems increase workcell complexity leading to the presence of more fault and failure modes. Given increasing manufacturing competitiveness, maximizing asset availability and maintaining desired quality and productivity targets have become essential. The National Institute of Standards and Technology (NIST) is developing measurement science (e.g., test methods, performance metrics, reference data sets) to monitor the degradation within a manufacturing workcell that includes a six-degree-of-freedom robot arm. Numerous components of the workcell influence the accuracy of the robot's tool center position. Identifying the component(s) responsible for process degradation prior to the process performing out of specification will provide manufacturers with advanced intelligence to maintain or maximize their performance targets and asset availability. NIST's research in robot workcell health promotes workcell component health characterization and develops methods and tools to verify and validate this approach. This paper presents the overall research plan and the efforts to date in developing appropriate test methods, identifying key sources of workcell degradation, and presenting baseline performance data that is leveraged for health assessment.

Keywords: condition monitoring, degradation, diagnostics, industrial robot systems, kinematics, manufacturing processes, manufacturing systems, prognostics, testbed, use cases, workcell.

INTRODUCTION

Robot systems and technologies are becoming more commonplace within manufacturing operations as they increase in capability, become easier to integrate, and become more affordable to industry. Systems are performing a range of operations including high precision tasks [1-4]. Advanced sensing, monitoring, and control technologies have enhanced robot systems and the workcells to which they contribute, to be more productive and efficient. As robot workcells are supporting a larger number of functions or the number of steps within a process presents greater variability, monitoring the workcell's manufacturing process and the health of the physical system becomes more critical. New and complex workcells present greater opportunities for faults and failures to emerge, especially faults and failures that have never been seen before. Monitoring faults and failures of robot workcells has also become more important as robots are used in more collaborative environments in closer proximity to human manufacturing partners; degraded or malfunctioning robots present safety concerns [5].

Maintaining the health of robot workcells is imperative to maximizing asset availability and maintaining minimum levels of productivity and process/part quality. One of the critical metrics that many manufacturers track is *Overall Equipment Effectiveness* (OEE). OEE can be tracked at multiple levels (e.g., at the factory level, assembly line, and equipment level) within a facility and is typically the multiplicative product of productivity, asset availability, and quality [6]. When a process is initiated for the first time, a baseline of performance and health is typically captured. Either that level of performance is acceptable or deemed insufficient where changes are made to increase one or more elements of OEE.

The OEE of a robot workcell is heavily influenced by the reliability (including repeatability) of the robot's positioning [7]. Many manufacturing robot workcells leverage one or more six degrees of freedom (6DOF) industrial robot arms to serve as positioners for end effectors (i.e., tools mounted to the flange at the end of the arm) to achieve a specific task. End effectors range from grippers used in material handling operations, to welding guns, or paint applicators used in very specific activities [2, 8, 9]. In some instances, the robot acts as the operation's macro-

manipulator, where the end effector serves as the micro-manipulator (e.g., a robot arm with a robotic gripper for in-hand manipulation). Other instances feature the robot as the sole positioner (e.g., a robot with an attached welding gun moving the welder into a specific position or along a trajectory). For these operations to maintain their OEE targets, the robot must be sufficiently reliable and repeatable.

Personnel at the U.S. Department of Commerce's National Institute of Standards and Technology (NIST) are developing the requisite measurement science to verify and validate monitoring, diagnostic, and prognostic capabilities to increase the reliability of manufacturing operations and minimize downtime (both planned and unplanned) [10, 11]. Numerous case studies [6, 12, 13] and the development of relevant manufacturing use cases [14] have been a driving force within this effort. These efforts have resulted in a portion of the project focusing on manufacturing operations of 6DOF industrial robot arms [7, 8, 15].

The goal of the article is to present the latest research plan that is propelling this effort, highlight what has been done to date, including recent accomplishments, and lay out the immediate next steps. One substantial goal of this effort is to provide industry with a low-cost, minimally invasive test method that can be applied within a manufacturing robot workcell to ascertain the health of the components that influence the kinematic chain that, in turn, influence the accuracy of the overall process and resultant product. The kinematic chain is the physical assembly of multiple rigid bodies that are connected to one another and constrained in specific degrees of freedom [7]. A known kinematic chain can be mathematically represented where equations can relate the position of one element to another element (in the chain). An error in a rigid body, or joint, in the kinematic chain can propagate through the rest of the chain creating a positional error.

In addition to disseminating this research in technical articles, it is expected that this effort will provide some of the technical basis for industry-driven standards [16-18]. Even though this specific research effort focuses on 6DOF industrial robot arms, the resultant methods can be adapted to accommodate industrial robot arms with greater or fewer degrees of freedom.

The remainder of this paper is organized as follows. The Background section presents prognostics and health management (PHM) and the critical role it is playing to advance manufacturing operations. The Workcell Research Focus section contains the bulk of this article and is divided into several sub-sections. The Research Motivation section includes the goals of the effort, the motivation, and the expected impact on industry. The Research Plan and Status section discusses the progress that has been made to date. The Current Efforts section discusses the active elements of the research. The Future Work and Conclusions section examines longer term efforts that are being planned to further this work and concludes the paper.

BACKGROUND

The field of Prognostics and Health Management (PHM) focuses on monitoring, diagnostic, and prognostic technologies to enhance maintenance and control strategies to maximize asset availability and maintain productivity and quality targets. With the emergence of technological innovations in manufacturing (i.e., the formation of Smart Manufacturing through greater connectivity of information technology and operations technology), PHM is quickly becoming a critical element within manufacturing operations. PHM has been applied by large manufacturers along with small to medium-sized manufacturers (SMMs) with differing degrees of success [6, 12, 13]. The manufacturing community has leveraged multiple PHM approaches (e.g., data-driven methods, physics-based models, and hybrid methods) to minimize their reactive maintenance activities and optimize their preventive and predictive maintenance efforts [19-22]. Some of the publicly-available PHM practices have been documented in a variety of standards documents [16, 17].

Smart manufacturing presents a complex environment for which PHM is being applied. Prior to smart manufacturing, manufacturing operations and systems were largely disconnected from one another; boundaries were very clear, and relationships among elements were relatively simplistic. The integration of advanced technologies and the connectivity of varying manufacturing processes and equipment across multiple physical locations makes it more challenging to effectively design, deploy, verify, and validate PHM. A manufacturing robot workcell can be considered a complex system of systems and therefore provides an appropriate use case for the application, and verification and validation (V&V), of PHM. A workcell is both an element of a much larger manufacturing operation, and an element that can be broken down into constituent sub-systems, components, sub-components, etc. Successful application of PHM within a workcell requires an understanding of the physical elements within that workcell and how they relate to one another. Maintenance activities are typically performed on physical elements (e.g., replacing a joint, lubricating gears) making it critical to understand how physical elements influence each other, not just in function, but also in health. The decomposition of physical elements of a robotic workcell into a representative hierarchy of elements provides a means of identifying boundaries that can drive maintenance tasks [7]. Furthermore, the physical hierarchy could then be integrated with informational and functional hierarchies to promote a greater understanding of the relationships among elements. This integration would also identify critical metrics and measures of workcell health, both in terms of process health and equipment health [23-25]. As system complexity increases, it becomes more critical to understand the inherent relationships to see how the state of mechanical degradation of physical elements impact process performance.

Recognizing that the robot is a critical element of the workcell, NIST is undertaking another research effort to use vision technology to capture the degradation in accuracy of a robot's tool center position (TCP) while the robot moves through

Weiss, Brian.
"DEVELOPING MEASUREMENT SCIENCE TO VERIFY AND VALIDATE THE IDENTIFICATION OF ROBOT WORKCELL DEGRADATION."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

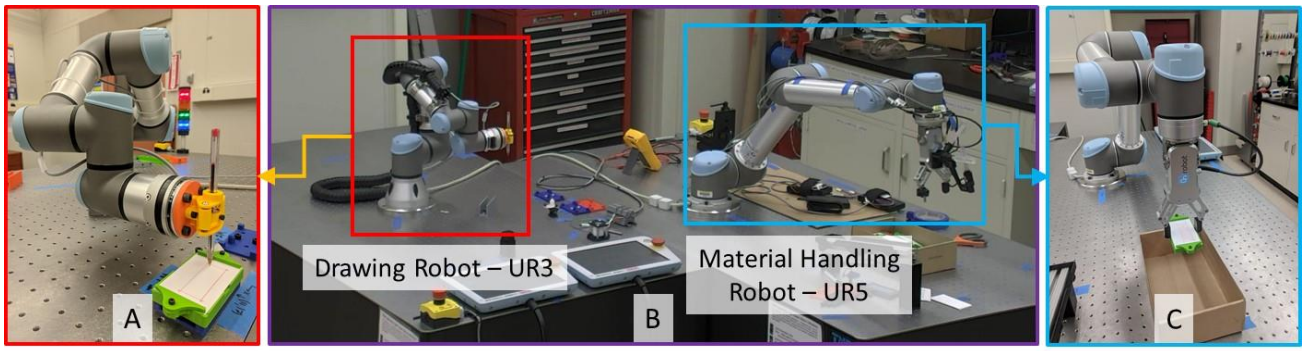


Figure 1: NIST MANUFACTURING PHM RESEARCH ROBOT WORKCELL (B); UR3 DRAWING ON A BUSINESS CARD AFFIXED TO A GREEN PLASTIC PART (BUSINESS CARD HOLDER) (A); UR5 PLACING A COMPLETED PART IN THE OUTPUT BIN (C)

a series of pre-programmed trajectories [26, 27]. The output of this work will provide an understanding of the overall robot arm health along with specific health intelligence of individual joints and constituent components (e.g., motors, encoders, gears). It is expected that this work will complement the efforts of the workcell-level research.

WORKCELL RESEARCH FOCUS

Similar to the definitions listed in the International Organization for Standardization (ISO) Standard 8373 for an Industrial Robot and an Industrial Robot System [28], this effort defines a robot as an automatically controlled, reprogrammable, multipurpose manipulator that is programmable in three or more axes; and the robot workcell as the one or more robots, end-effector(s), and any machinery, supporting automation, external axes, sensors, work fixtures, etc. necessary to accomplish a specific task.

The robot workcell use case leveraged in this research effort is defined as two, 6DOF industrial robot arms working together (shown in Figure 1B with the purple border) to complete a specific task. One robot, a Universal Robots UR5 (shown in Figure 1B and Figure 1C with blue borders), is tasked with performing material handling operations. This robot has a movable gripper mounted to its tool flange where the robot's controller commands the gripper to open and close. The other robot's (a Universal Robots UR3 shown in Figure 1A and Figure 1B with red borders) end-effector is a custom-built holder that contains a pen to support the robot's ability to draw on a surface. The two robot controllers receive higher-level commands from a supervisory Programmable Logic Controller (PLC), and work together to draw a specified design on a business card that is clamped to a plastic part (a green part is shown in Figure 1A). The relevance of this specific use case to that of actual manufacturing operations is 1) the material handling operations of moving a part from an input tray to a work fixture, and then from the work fixture to an output bin, is comparable to part pick and placement common in manufacturing operations; 2) the drawing robot is performing a task comparable to manufacturing path planning operations (e.g., welding or adhesive application) where the robot serves as the lone manipulator for a precision tool and must move along a specific trajectory at a specific speed and accuracy (e.g., vibrations of the end-effector must be within specified tolerances) as it modifies the part; and 3) the overall workcell could be compared to a machine tending operation where a material handling robot places a part within the work volume of a machine tool or an additive manufacturing tool. In this instance, the drawing robot is modifying the part through an additive process. Drawing on a business card, as opposed to drawing directly on the plastic part, allows for quick and cost-effective replacement of the modified part, as opposed to fabricating additional plastic parts.

The workcell's manufacturing process begins with a part being picked up by the material handling robot from the input tray (see Figure 2A and Figure 2B). The material handling robot places (see Figure 2C) the part on one of two fixtures within range of the path planning robot. Once the part is placed on a fixture, the path planning robot begins drawing the target design on the business card within the part (see Figure 3D). While the UR3 is completing this task, the UR5 is moving another part (see Figure 3E and Figure 3F) to the second work fixture, if the work fixture is available and another part has been ordered for drawing. When the path planning robot has completed this task, the material handling robot removes the part from its fixture and deposits it in an output box (see Figure 4G). It is important to note that there are two fixtures within the work volume that can be used by both robots enabling two parts to be 'in process' at the same time. After the UR3 is done drawing on the last part (see Figure 4H), the UR5 removes it from its fixture (see Figure 4I)and moves it to the output bin to complete the production run.

The remainder of this section presents NIST's specific research efforts including the motivation behind this research, the overall research plan, the status of the work, and the current efforts.

Research Motivation

NIST's research plan is built upon addressing the following questions, based upon the Heilmeier catechism (sometimes known as the Heilmeier questions) [29], to articulate the appropriateness and value of the research.

- *What is the problem we are trying to solve? Why is it important?*

Weiss, Brian.
"DEVELOPING MEASUREMENT SCIENCE TO VERIFY AND VALIDATE THE IDENTIFICATION OF ROBOT WORKCELL DEGRADATION."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

The goal is to offer a solution to enable the manufacturing community to verify and validate emerging technologies that monitor, diagnose, and predict the health of robot workcells. As noted earlier, the workcell extends beyond the robot arm including end-effectors, work fixtures, and parts. This involves providing the means to determine the overall health of a robot workcell along with the source of degradations prior to the degradations lowering part and/or process quality, and productivity, out of specification. This would provide the manufacturing community with a test procedure and corresponding performance metrics that would test specific elements (e.g., robot arm, end-effector) of the workcell to determine the degradation, if any, of each element with respect to its influence on process/robot accuracy. Without this capability, manufacturers would be unaware of any degradations within their workcell until it is observed in lower productivity or quality measures, or are observed when a physical component fails.

- *What are we trying to accomplish?*

  NIST personnel are trying to develop the requisite measurement science, including test methods, performance metrics, and reference datasets, and contribute to standards to verify and validate emergent technologies that monitor, diagnose, and predict workcell health by examining the physical elements that impact process and robot accuracy. To that end, NIST is conducting multiple research efforts including 1) producing reference datasets from its representative manufacturing workcell to support PHM algorithm development, verification, and validation and 2) developing a minimally-invasive test method to serve as a low-cost solution that will assess the health of a robot workcell and identify the source(s) of degradations prior to their impact on performance targets.

- *How does this get done at present? What are the limitations of current approaches?*

  Current workcell degradations are typically detected when either part and/or process quality (e.g., accuracy of a process), or productivity, has fallen below target thresholds. If workcell degradation is detected in advance, it involves one or more of: costly test equipment, time-consuming processes, and/or test equipment disruptive to the workcell configuration. For example, some manufacturers, especially those that require high precision performance from their robot systems, will use laser-based systems to measure changes in a robot's accuracy. This information is typically used to recalibrate the robot and/or identify changes in accuracy that indicate degraded health.

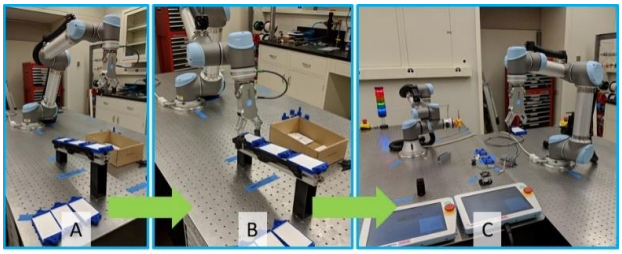


Figure 2. UR5 APPROACHING THE INPUT TRAY (A), PICKING UP A PART (B), AND MOVING THE PART TO AN OPEN FIXTURE (C).

- *Why should NIST do it?*

  The manufacturing community presents a diversity of robot workcells that feature a variety of configurations, robot manufacturers, end effectors, and supporting automation (e.g., linear rails, conveyors). NIST is an independent, neutral third party relative to the manufacturing industry. NIST has a strong history of developing device-agnostic test methods to objectively measure the capabilities of differing implementations [30-32]. NIST's research focus within manufacturing and robotics makes this problem relevant to the current mission.

- *What is new about this approach? Why do you think you can be successful at this time?*

  The approach focuses on assessing the kinematic chain that directly impacts the motion that is performed on the part or with the part. The novelty is the systematic and relatively simplistic test method that measures the repeatability of specific elements along the kinematic chain to determine if

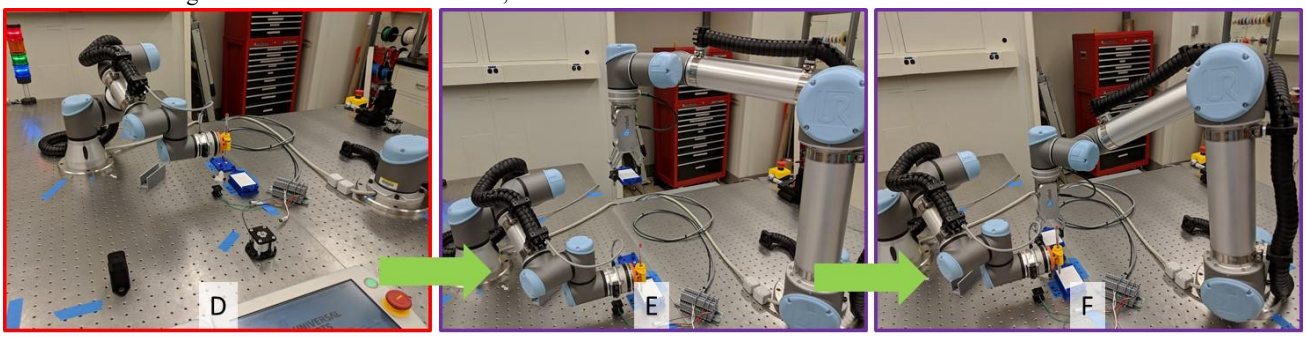


Figure 3. UR3 APPROACHING PART 01 ON FIXTURE 1 (D), UR3 DRAWING ON PART 01 AND UR5 MOVING PART 02 TO FIXTURE 2 (E), UR5 PLACING PART 02 ON FIXTURE 2 (F)

Weiss, Brian.
"DEVELOPING MEASUREMENT SCIENCE TO VERIFY AND VALIDATE THE IDENTIFICATION OF ROBOT WORKCELL DEGRADATION."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

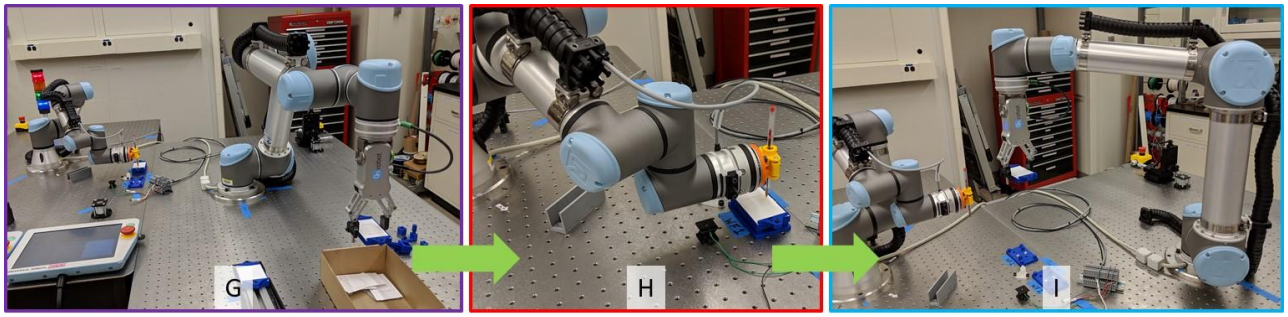


Figure 4. UR5 MOVING COMPLETED PART 01 TO OUTPUT BIN (G), UR3 DRAWING ON PART 02 IN FIXTURE 2 (H), UR5 MOVING PART 02 FROM FIXTURE 2 TO THE OUTPUT BIN (I)

they have deviated from their baseline specification. This research includes the development of a low-cost, innovative sensor that supports a minimally-invasive testing process which can be executed before or after the workcell's manufacturing operations with minimal impact to the cell's productivity. There is a high probability of success for this effort given that the test method's constituent sensor can be replicated at minimal cost and flexibly deployed within a target workcell. This approach is discussed in greater detail in the Current Efforts subsection in this paper.

- *Who cares?*

  Anyone in the manufacturing industry that manufactures robots, integrates robots into manufacturing operations, or uses robots in their factory will care about this effort. These stakeholders can all benefit from the successful development and dissemination of this effort. These groups should care about this work because process and equipment downtime due to maintenance, especially unplanned downtime due to faults or failures, can be a substantial cost to an organization.

- *What impact will success have? How will it be measured?*

  The successful execution and dissemination of this work is expected to have tremendous impact on the manufacturing community. The manufacturing community will gain a reliable and consistent method that they can integrate into many of their robot workcell operations including those that feature robots performing material handling or path planning operations. Technology integrators will gain a test method that they can build into the robot workcells that they supply their customers. Likewise, robot manufacturers will have a greater awareness of how their products are tested in-situ which may drive them to build the robot-based test method components into their robots while the robots are being fabricated. Measuring this impact will include capturing the number of manufacturers that integrate this method into their robot-based workcells and the number of active robot workcells that incorporate this method. Similarly, the number of integrators that choose to offer this test method as a PHM option and the number of workcells they output with this method will be captured.

Research Plan and Status

This research plan follows the path of:

1. Conduct case studies – Completed with the generation of a comprehensive workshop report identifying numerous roadmap action plans of key measurement science challenges in the PHM field [10]. Action plans described in this report that motivate the robot workcell research include: Advanced Sensors for PHM in Smart Manufacturing, Identification of PHM Performance Metrics, and Failure Data for Prognostics and Diagnostics.
2. Identify an appropriate use case(s) – Completed with the determination of the two-robot workcell use case. This configuration can also be considered an abstraction of a robot and machine tool workcell [8, 14]. Part of this effort has also included identifying the different degradation modes of the workcell [7].
3. Identify critical performance metrics – In Process. This has begun with determining that the workcell's kinematic chain will be monitored regarding its influence on the accuracy of the robot's TCP and the accuracy of the part's movement within environment [7]. Similarly, process-level metrics

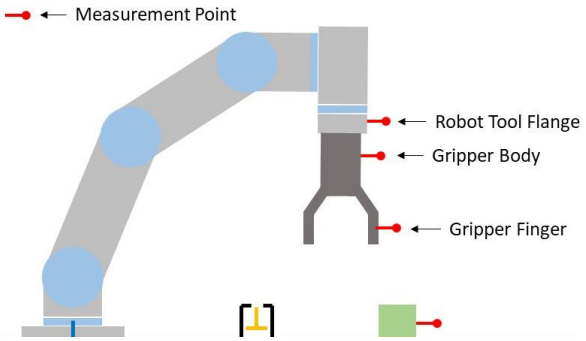


Figure 5. VISUAL REPRESENTATION OF TEST POINTS ALONG THE KINEMATIC CHAIN OF THE ROBOT AND OTHER WORKCELL ELEMENTS

Weiss, Brian.
"DEVELOPING MEASUREMENT SCIENCE TO VERIFY AND VALIDATE THE IDENTIFICATION OF ROBOT WORKCELL DEGRADATION."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

have been identified that that can inform about the workcell's productivity during operation of the use case [15]. Additional metrics are still being explored for inclusion.

4. Develop test methods – In Process. A test method that monitors the health of the kinematic chain at various points [along the chain] has been developed [7]. Figure 5 presents a visual representation of multiple test points along the kinematic chain of the robot, the gripper, and the part that can all uniquely influence part and process quality. The test method is being verified at NIST. Likewise, technology integrators and manufacturers are being engaged to validate the implementation within actual manufacturing operations. The verification and validation efforts will be discussed further in the Current Efforts section.
5. Capture reference dataset(s) – In Process. This will be discussed in detail in the Current Efforts section.
6. Contribute technical basis to the standards community – Not yet started. Test method verification and validation must occur prior to this research being introduced into the standards community. In support of the NIST's PHM research efforts, NIST personnel have been a driving force in the creation of a newly formed American Society of Mechanical Engineers (ASME) subcommittee on Advanced Monitoring, Diagnostics, and Prognostics for Manufacturing Operations [33]. The vision is that the technical basis of the robot workcell research will be integrated into guideline documents developed by this subcommittee when the research has sufficiently matured.

Data Collection, Verification, and Validation

Current efforts are focused on capturing reference datasets within the representative manufacturing workcell, verification of the novel sensor that has been developed to detect degradation in the kinematic chain, and validation of the kinematic chain test method within industry.

The kinematic chain test method and the corresponding discrete positioning sensor have been developed at NIST. As discussed in [7], this test method relies upon the inspection of the workcell's kinematic chain to identify and track degradation of workcell elements. This is accomplished by measuring the positioning repeatability of critical points (shown in Figure 5) along the kinematic chain. The test method is paired with a custom-built sensor that indicates discrete measurements of whether an element along the kinematic chain is maintaining its accuracy and therefore, repeatability. This technology is described as a "Position Verification Sensor with Discrete Output" (U.S. Provisional patent application serial number 62/732,059) and is shown in Figure 6. The overall test method can be executed with other sensing technology; the development of this new sensor was motivated by providing the community with a relatively low-cost solution that can be mass produced and readily integrated within many existing workcells. The expectation is that the sensor will be produced at a cost of between $50 to $100 USD (less if mass-produced). The sensor and corresponding test approach are likely to be an economical alternative to reactive maintenance. Capturing data from the sensor at specific time horizons or before/after certain activities can influence scheduling of maintenance activities in an effort to minimize workcell downtime and maintenance costs.

The sensor provides feedback when a cylindrical pin is vertically inserted into the top of the sensor within the manufactured tolerances such that only the inner button (shown within the green inner circle in Figure 6) is depressed, and the outer surface (shown within the red outer surface perimeter circle in Figure 6) is not touched. The implementation of the kinematic chain test method calls for one or more sensors to be placed within the work volume of the robot(s) within the workcell. The kinematic chain test method is engaged after a specified amount of production cycles or is directly forced by the operator through the PLC during certain windows (e.g., at the start of a shift, at the conclusion of the work day, or as part of preventive/routine maintenance). Two sensors are currently deployed in NIST's representative manufacturing workcell – an early prototype that is within the reach of the UR3, and the current prototype (shown in Figure 6) that is within the reach of the UR5.

For the NIST use case, test method execution begins with the UR3 moving its pen tip and attempting to press the inner button of the sensor within its reach. Next, the UR3 tests attempts to insert the pin (attached just above its tool flange) into the sensor. Success or failure is noted. The kinematic chain of the UR5 is then tested. This subprocess begins with testing the robot arm. This is shown in both Figure 7A and Figure 7B. After the robot arm is tested, the gripper body is tested, and the gripper fingers (while open) are tested. These activities are shown in Figure 7C and Figure 7D. Next, the gripper fingers (while closed) are tested. Lastly, the robot picks up a test part, that contains a vertical pin, and manipulates the test part with the sensor to yield a pass or fail result. The results of this test method produce a combination of pass and/or fail results regarding the

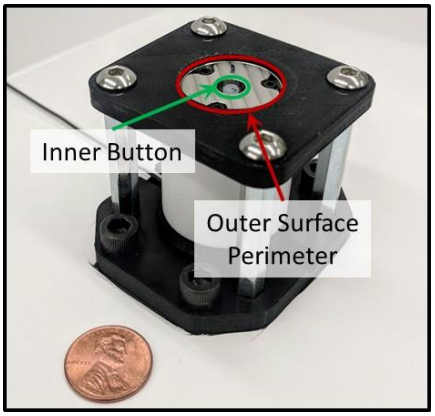


Figure 6. "POSITION VERIFICATION SENSOR WITH DISCRETE OUTPUT" – U.S. PROVISIONAL PATENT APPLICATION SERIAL NUMBER 62/732,059

Weiss, Brian.
"DEVELOPING MEASUREMENT SCIENCE TO VERIFY AND VALIDATE THE IDENTIFICATION OF ROBOT WORKCELL DEGRADATION."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

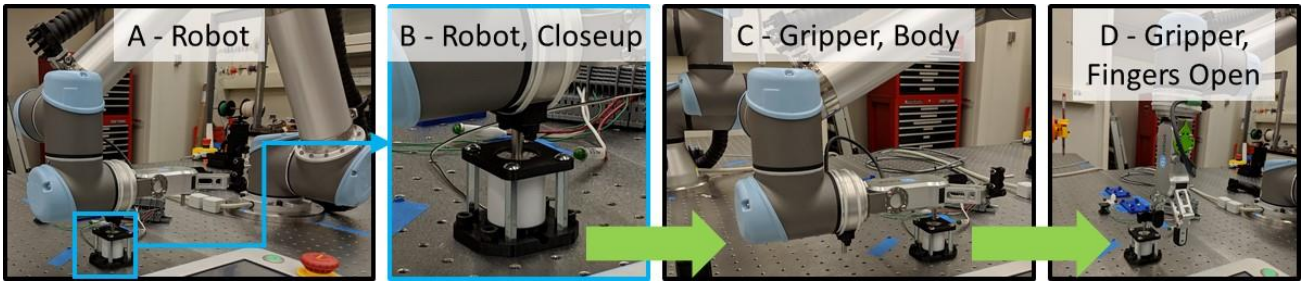


Figure 7. SUBSET OF TEST POINTS OF THE KINEMATIC TEST METHODOLOGY APPLIED TO THE NIST UR5

elements that were tested. Analysis of the results can highlight which, if any, of the elements in the kinematic chain are out of the tested specification. This information can be used to further troubleshoot specific elements of the workcell, perform maintenance on one or more elements or recalibrate an element of the workcell until the necessary maintenance can be performed. Feedback from this test method can also be correlated with process level and robot controller data to further isolate sources of degradation.

There is value to both NIST researchers and the manufacturing community in capturing reference datasets and making them publicly available. Reference datasets afford NIST the opportunity to develop, verify, and validate its methods and tools on representative manufacturing data from its robot workcell. Data captured in this environment presents realistic variation similar to data collected from an actual manufacturing facility. The manufacturing community benefits from accessing these NIST reference datasets since these datasets are freely available (i.e., no cost), will be annotated, contain more context than is typical for data captured in real manufacturing operations, provide data to support innovative technology development, and offer objective data to promote independent technology assessments. A series of data collections are planned with the first data collection being complete. These specific collections are driven by the need to capture a baseline of health and performance under several reasonable operating scenarios; followed by collections under various degradation conditions (either real or simulated). The steps in the series are:

- Baseline operation with plastic parts (28 g per part), no simulated degradation modes, all part geometrics are within tolerances, fixture geometry is within tolerances, gripper fingers are within tolerances. 60 parts are run.
- Baseline operation with **heavier parts (>> 28 g)** (e.g., steel or aluminum – exact mass is to be determined), no simulated degradation modes, all part geometrics are within tolerances, fixture geometry is within tolerances, gripper fingers are within tolerances.
- Operation with plastic parts, **with simulated backlash** (i.e., backlash can be simulated at each of the six joints of each robot arm. The exact joint(s) and robots that will present the backlash are still to be determined), all part geometrics are within tolerances, fixture geometry is within tolerances, gripper fingers are within tolerances.
- Operation with plastic parts, **with simulated slip** (i.e., slip can be simulated at each of the six joints of each robot arm. The exact joint(s) and robots that will present the slip are still to be determined), all part geometrics are within tolerances, fixture geometry is within tolerances, gripper fingers are within tolerances.
- Operation with plastic parts, no simulated degradation modes, **part degradation with respect to its interface on the fixture** (the number of parts to be degraded in the experiment is still to be determined), fixture geometry is within tolerances, gripper fingers are within tolerances.
- Operation with plastic parts, no simulated degradation modes, all part geometrics are within tolerances, **fixture geometry is degraded with respect to its interface with parts, gripper fingers are within tolerances**.
- Operation with plastic parts, no simulated degradation modes, all part geometrics are within tolerances, fixture geometry is within tolerances, **geometry of gripper fingers where they contact the part is degraded** (exact degradation and if it will occur on one or both fingers are still to be determined).
- Operation with a **combination of degradations**

The first dataset was captured from the workcell during the representative manufacturing operation where 60 parts were 'processed.' This involved six unique business card holders, numbered 1 – 6, being cycled through the workcell with blank business cards justified to the bottom left of the holder. Prior to the execution of the 60 runs, the robots cycled through the motions of processing 30 parts, yet no parts were fed to the robot during this time (this was done as a purposeful warmup). The input tray was loaded with an initial three parts (i.e., business card holders 1 through 3) shown in Figure 2B. PLC data monitoring and collection was turned on for process data and robot controller-level data for both the UR3 and UR5. Robot data was also directly captured from the robots (in addition to being captured through the PLC). The reason robot data is captured from two separate sources is because capturing it directly from the robots' controllers provides high resolution data; capturing

Weiss, Brian.
"DEVELOPING MEASUREMENT SCIENCE TO VERIFY AND VALIDATE THE IDENTIFICATION OF ROBOT WORKCELL DEGRADATION."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

lower resolution robot controller data through the PLC adds a greater measure of assurance regarding time synchronization between the process and robot controller data feeds since the PLC time is captured in all data files. Additionally, data is collected from the OptoForce Force/Torque sensor that is mounted between the tool flange and pen holder on the UR3. The 60 runs were completed in approximately 48 minutes. Besides capturing the noted data files, the robot script files, configuration files, and log files were captured from the UR3 and UR5 in case any operational faults were discovered while reviewing the data.

Presently, the dataset is under examination where the conclusions will be presented in a future article. Likewise, the lessons learned from this dataset will inform on the expected next data collection that will feature increasing the weight of the parts.

In parallel with capturing reference datasets on the workcell's operations under varying conditions, it is important to verify the discrete sensor. To date, manual verification of the sensor has been done using a hand-driven 3DOF linear stage (shown in Figure 8). Automated verification is planned with motorized drives that will increase the efficiency of this activity and test in a random pattern (as opposed to the very deterministic pattern used during manual verification).

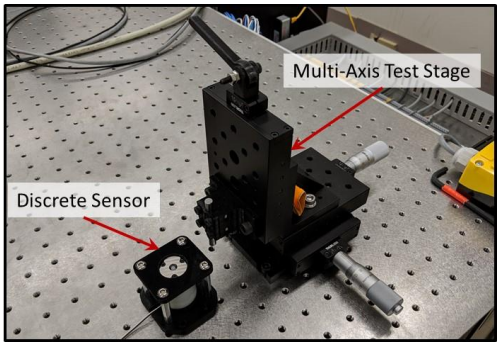


Figure 8. DISCRETE SENSOR AND TEST STAGE

The other active effort focuses on the validation of the kinematic chain test method. This involves the determination of the appropriateness of the test method within relevant manufacturing environments. Collaborations are being explored with external partners including a robotics distributor, technology integrators, and manufacturers. The expectation is that the kinematic chain test method will be piloted in a manufacturing environment or integrated into a developmental cell.

Integrating one or more developmental sensors in a functional manufacturing workcell along with executing the test method should offer valuable feedback to NIST researchers on the validity of the test method, the performance metrics being captured, and the viability of the sensor. Likewise, manufacturers will have the opportunity to discover the presence of any health degradations across their robot's kinematic chain and determine where, along the kinematic chain, degradations are present. Given the developmental status of the sensor, the sensor has yet to be ruggedized for extensive use in an actual environment. The expectation is that the sensor's deployment will be limited to relatively clean workcells (e.g., workcells with minimal to no usage of fluids or lubricants, and workcells that do not output metal chips).

FUTURE WORK AND CONCLUSIONS

A strong foundation is established to conduct research in developing the appropriate measurement science to verify and validate robot workcell PHM technology. The manufacturing community has a need for independently-developed test methods, reference datasets, and guidelines to assess and advance the state of the art in monitoring, diagnostic, and prognostic technologies for manufacturing robot workcells. To date, case studies have shown a need for the measurement science, a use case has been articulated, and a test bed has been constructed. Performance metrics have been identified and a test method has been produced which are both still being iterated upon. More recently, a reference dataset has been captured with additional datasets being planned; manual verification of the sensor has been completed with a more comprehensive automated verification being planned; and validation of the test method and sensor are being explored. As the workcell-level test methods mature, the feedback from these test methods will feed into the robotic-level testing which heavily factors in the robot's kinematic model and aims to identify specific joint errors. Another intersection of the robot-level and workcell-level effort will be that of simulating slip and backlash of specific joints at the robot level (this is already mentioned in the workcell-level research and is expected to be done concurrently at the robot-level). Ultimately, this measurement science will be transitioned into standards or guidelines to further disseminate this work and promote best practices of assessing robot workcell health.

ACKNOWLEDGEMENTS

The author would like to acknowledge the contributions of Alexander Klinger, a Robotics Engineer at Titan Robotics. Mr. Klinger was a former member of the NIST research team and is a co-inventor of the discrete sensor.

NIST DISCLAIMER

The views and opinions expressed herein do not necessarily state or reflect those of NIST. Certain commercial entities, equipment, or materials may be identified in this document to illustrate a point or concept. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

REFERENCES

[1] Ahmad, R., and Plapper, P., 2016, "Safe and Automated Assembly Process using Vision assisted Robot Manipulator," Research and Innovation in Manufacturing: Key Enabling Technologies for the Factories of the Future - Proceedings of the 48th Cirp Conference on Manufacturing Systems, 41, pp. 771-776.

Weiss, Brian.
"DEVELOPING MEASUREMENT SCIENCE TO VERIFY AND VALIDATE THE IDENTIFICATION OF ROBOT WORKCELL DEGRADATION."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

[2] Chen, H., Fuhlbrigge, T., and Li, X., "Automated industrial robot path planning for spray painting process: a review," Proc. Automation Science and Engineering, 2008. CASE 2008. IEEE International Conference on, IEEE, pp. 522-527.
[3] Marvel, J. A., and Norcross, R., 2017, "Implementing Speed and Separation Monitoring in Collaborative Robot Workcells," Robot Comput Integr Manuf, 44, pp. 144-155.
[4] Marvel, J. A., Saidi, K., Eastman, R., Hong, T., Cheok, G., and Messina, E., "Technology readiness levels for randomized bin picking," Proc. Proceedings of the Workshop on Performance Metrics for Intelligent Systems, ACM, pp. 109-113.
[5] Marvel, J. A., Falco, J., and Marstio, I., 2015, "Characterizing Task-Based Human–Robot Collaboration Safety in Manufacturing," Systems, Man, and Cybernetics: Systems, IEEE Transactions on, 45(2), pp. 260-275.
[6] Jin, X., Siegel, D., Weiss, B. A., Gamel, E., Wang, W., Lee, J., and Ni, J., 2016, "The present status and future growth of maintenance in US manufacturing: results from a pilot survey," Manuf Rev (Les Ulis), 3, p. 10.
[7] Klinger, A., and Weiss, B. A., 2018, "Examining Workcell Kinematic Chains to Identify Sources of Positioning Degradation," Annual Conference of the PHM SocietyPhiladelphia, Pennsylvania, p. 9.
[8] Weiss, B. A., and Klinger, A. S., "Identification of Industrial Robot Arm Work Cell Use Cases and a Test Bed to Promote Monitoring, Diagnostic, and Prognostic Technologies," Proc. 2017 Annual Conference of the Prognostics and Health Management (PHM) Society, PHM Society, p. 9.
[9] Agheli, M., Qu, L., and Nestinger, S. S., 2014, "SHeRo: Scalable hexapod robot for maintenance, repair, and operations," Robotics and Computer-Integrated Manufacturing, 30(5), pp. 478-488.
[10] Pellegrino, J., Justiniano, M., Raghunathan, A., and Weiss, B. A., 2016, "Measurement Science Roadmap for Prognostics and Health Management for Smart Manufacturing Systems," NIST Advanced Manufacturing Seriess (AMS).
[11] Weiss, B. A., Vogl, G. W., Helu, M., Qiao, G., Pellegrino, J., Justiniano, M., and Raghunathan, A., 2015, "Measurement Science for Prognostics and Health Management for Smart Manufacturing Systems: Key Findings from a Roadmapping Workshop," Annual Conference of the Prognostics and Health Management Society 2015, P. Society, ed., PHM Society, Coronado, CA, p. 11.
[12] Jin, X., Weiss, B. A., Siegel, D., and Lee, J., 2016, "Present Status and Future Growth of Advanced Maintenance Technology and Strategy in US Manufacturing," Int J Progn Health Manag, 7(Spec Iss on Smart Manufacturing PHM), p. 012.
[13] Helu, M., and Weiss, B. A., "The current state of sensing, health management, and control for small-to-medium-szed manufacturers," Proc. ASME 2016 Manufacturing Science and Engineering Conference, MSEC2016.
[14] Weiss, B. A., Helu, M., Vogl, G. W., and Qiao, G., 2016, "Use Case Development to Advanced Monitoring, Diagnostics, and Prognostics in Manufacturing Operations," IMS2016 - Intelligent Manufacturing SystemsAustin, Texas, p. 6.
[15] Klinger, A. S., and Weiss, B. A., 2018, "Robotic Work Cell Test Bed to Support Measurement Science for PHM," 2018 ASME Manufacturing Science and Engineering Conference (MSEC), American Society of Mechanical Engineers (ASME), College Station, Texas.
[16] Vogl, G. W., Weiss, B. A., and Donmez, M. A., 2014, "Standards Related to Prognostics and Health Management (PHM) for Manufacturing," No. NISTIR 8012, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA.
[17] Vogl, G. W., Weiss, B. A., and Donmez, M. A., "Standards for prognostics and health management (PHM) techniques within manufacturing operations," Proc. Annual Conference of the Prognostics and Health Management Society 2014.
[18] Weiss, B. A., Alonzo, D., and Weinman, S. D., 2017, "Summary Report on a Workshop on Advanced Monitoring, Diagnostics, and Prognostics for Manufacturing Operations."
[19] Lee, J., Lapira, E., Bagheri, B., and Kao, H.-a., 2013, "Recent advances and trends in predictive manufacturing systems in big data environment," Manufacturing Letters, 1(1), pp. 38-41.
[20] Lee, J., Ni, J., Djurdjanovic, D., Qiu, H., and Liao, H. T., 2006, "Intelligent prognostics tools and e-maintenance," Computers in Industry, 57(6), pp. 476-489.
[21] Peng, Y., Dong, M., and Zuo, M. J., 2010, "Current status of machine prognostics in condition-based maintenance: a review," International Journal of Advanced Manufacturing Technology, 50(1-4), pp. 297-313.
[22] Vogl, G. W., Weiss, B. A., and Helu, M., 2016, "A review of diagnostic and prognostic capabilities and best practices for manufacturing," Journal of Intelligent Manufacturing.
[23] Sharp, M., and Weiss, B. A., 2018, "Hierarchical Modeling of a Manufacturing Work Cell to Promote Contextualized PHM Information Across Multiple Levels," Manuf Lett, 15(A), pp. 46-49.
[24] Weiss, B. A., and Qiao, G., 2017, "Hierarchical Decomposition of a Manufacturing Work Cell to Promote Monitoring, Diagnostics, and Prognostics," ASME 2017 International Manufacturing Science and Engineering Conference (MSEC2017), ASME, Los Angeles, California, p. 11.
[25] Weiss, B. A., Sharp, M., and Klinger, A., 2018, "Developing a hierarchical decomposition methodology to increase manufacturing process and equipment health awareness," Journal of Manufacturing Systems, 48, pp. 96-107.
[26] Qiao, G., Schlenoff, C. I., and Weiss, B. A., 2017, "Quick Positional Health Assessment for Industrial Robot Prognostics and Health Management (PHM)," IEEE International Conference on Robotics and Automation 2017Singapore, p. 6.
[27] Qiao, G., and Weiss, B. A., 2017, "Accuracy Degradation Analysis for Industrial Robot Systems," ASME International Manufacturing Science and Engineering Conference, ASME, Los Angeles, California, p. 9.
[28] International Organization for Standardization, 2012, "ISO 8373: Robots and Robotic Devices - Vocabulary," ISO, Switzerland, p. 48.

[29] Logar, N., 2009, "Towards a culture of application: science and decision making at the National Institute of Standards & Technology," Minerva, 47(4), pp. 345-366.
[30] Weiss, B. A., Schlenoff, C., Sanders, G. A., Steves, M. P., Condon, S. L., Phillips, J., and Parvaz, D., "Performance Evaluation of Speech Translation Systems."
[31] Weiss, B. A., Schlenoff, C., Shneier, M., and Virts, A., "Technology evaluations and performance metrics for soldier-worn sensors for assist."
[32] Weiss, B. A., and Menzel, M., 2010, "Development of domain-specific scenarios for training and evaluation of two-way, free form, spoken language translation devices," National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA.
[33] ASME, 2018, "Codes & Standards - PHM Subcommittee on Monitoring, Diagnostic, and Prognostic for Manufacturing Operations," https://cstools.asme.org/csconnect/CommitteePages.cfm?Committee=102342234.

Weiss, Brian.
"DEVELOPING MEASUREMENT SCIENCE TO VERIFY AND VALIDATE THE IDENTIFICATION OF ROBOT WORKCELL DEGRADATION."
Paper presented at Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, United States. June 10, 2019 - June 14, 2019.

Proceedings of the ASME 2019 14th International Manufacturing Science and Engineering Conference
MSEC2019
June 10-14, 2019, Erie, PA, USA

MSEC2019-2911

BEARING METRICS FOR HEALTH MONITORING OF MACHINE TOOL LINEAR AXES

Gregory W. Vogl[1]
National Institute of Standards and Technology
Gaithersburg, MD, USA

Brian C. Galfond
Catholic University of America
Washington, DC, USA

N. Jordan Jameson
National Institute of Standards and Technology
Gaithersburg, MD, USA

ABSTRACT

Diagnostics and prognostics of rotating machinery ball bearings is quite mature with an abundance of available methods and algorithms. However, extending these algorithms to other ball bearing applications is challenging and may not yield usable results. This work used a linear axis to study the ability of an inertial measurement unit (IMU), along with nine signal features, to measure changes in geometric error motions due to induced faults on the recirculating ball bearings of two carriage trucks. The IMU data was analyzed with the nine features used for rotating machinery systems, including root-mean-square, standard deviation, and kurtosis. For each stage of degradation, the statistical population and median value of each feature were compared against the population and median for no degradation, to monitor feature changes due to ball damage. Correlation analyses revealed an ability of the standard deviation feature to detect statistically significant changes as small as 0.05 micrometers or 0.5 microradians, corresponding to a total damaged surface area of truck balls of less than 0.1 percent.

Keywords: machine tool, smart manufacturing, linear axis, ball bearing, wear, degradation, diagnostics

1. INTRODUCTION

Linear axes are vital components in manufacturing, existing within machine tools to move cutting tools and workpieces to their desired positions for part production [1]. Figure 1A shows a typical linear axis, composed of four trucks (also called "linear motion guides") that constrain a carriage to move along two rails (or "guideways"). As shown in Figure 1B, in addition to the commanded motion along the X-direction, the actual motion has three translational errors and three angular errors: one positioning error motion (E_{XX}), two straightness error motions (E_{YX} and E_{ZX}), and three angular error motions (E_{AX}, E_{BX}, and E_{CX}). Typically, a machine tool includes multiple axes that degrade with use, leading to changes in geometric motion errors that affect the quality of the machined parts.

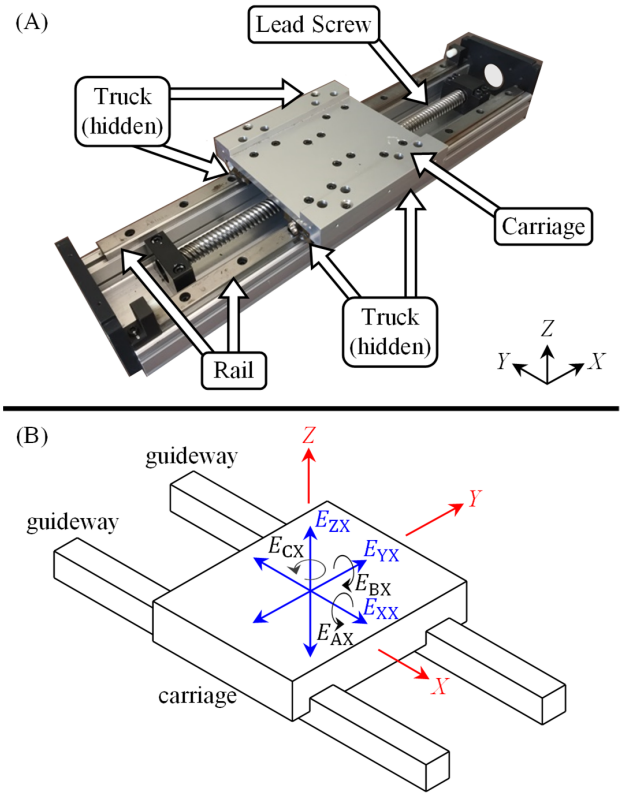


FIGURE 1: (A) LINEAR AXIS AND (B) GEOMETRIC ERROR MOTIONS.

[1] Contact author: gvogl@nist.gov

Vogl, Gregory; Galfond, Brian; Jameson, Jordan.
"BEARING METRICS FOR HEALTH MONITORING OF MACHINE TOOL LINEAR AXES."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

Ideally, once a linear axis is operational on a machine tool, all error motions are zero; but in practice, nonzero errors exist that contribute to the errors on the workpiece. These error motions tend to worsen with machine usage, aging, and crashes, since abrasion and adhesion between parts in linear axes causes material fatigue, pitting, cracking, and wear. This damage can result in faults developing in linear axis components, such as the rails, rolling element bearings, and/or ball screw [2, 3]. If not properly mitigated, these faults will grow to affect the quality of parts produced, leading to parts becoming out of tolerance and/or machine failure [4]. As demands for versatility and batch volume increase for manufacturing processes, machines are experiencing higher production loads, and as a result, the potential for faults and failures is becoming more common. Hence, machine tool monitoring and maintenance rules are needed to mitigate this accumulation of degradation and minimize the costs imposed by imperfect production and scrapped parts.

Mature methods exist for the fault detection and diagnostics of error motions, but they are manual, time consuming, and often cost prohibitive. The state-of-the-art instruments for linear axis error measurement (the basis for diagnostics) are explained in the International Organization for Standardization (ISO) 230-1 [5]: straightedge and linear displacement sensor, microscope and taut wire, alignment telescope, alignment laser, and laser straightness interferometer. These time-consuming measurements require a shutdown of the machine with a typical setup change and thus cannot provide in-situ diagnostics [6].

To enable proactive (not reactive) maintenance, manufacturers need automated methods for diagnosing machine tool linear axes without halting production. In 2010, Teti *et al*. [7] identified that intelligent sensor-based systems and advanced signal data processing need to be further developed to help decrease machine downtime and increase productivity, product quality, and knowledge of manufacturing processes. One possible advance lies in the use of an inertial measurement unit (IMU) consisting of a three degree-of-freedom (DOF) accelerometer and a three DOF rate gyroscope [8-10], as shown in Figure 2. Data from the IMU can be used to detect changes in the positioning, straightness, and angular error motions. IMU measurements can be made quickly and with little intrusion into the operation of the machine, resulting in data that provides insight into the condition of the linear axis. It has been shown to be effective at detecting rail degradation to similar levels of accuracy delivered by a laser interferometer [8].

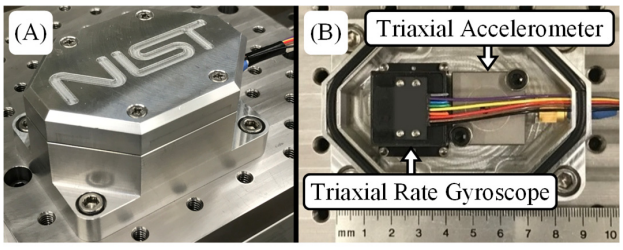


FIGURE 2: (A) ISOMETRIC VIEW OF IMU AND (B) TOP VIEW OF IMU WITHOUT ITS COVER.

However, a challenge remains in verifying the IMU's capabilities for detecting degradation in the truck bearings of a linear axis. One complication is the multitude of rolling element bearings (typically hundreds of balls) within the bearing system, which causes a low signal-to-noise ratio (that can be less than 1) in the health monitoring data. The convolution of numerous ball imperfections affects the geometric error motions of the carriage, making it difficult to isolate small influences due to damage on a single ball. Accordingly, an experiment was designed to examine the sensitivity of the IMU-based error motions to artificially-induced damage on the truck bearings.

2. EXPERIMENTAL SETUP

Figure 3 shows the linear axis testbed used for data collection in this study. A lead screw rotates via a DC motor to move a carriage nominally parallel with the X-axis (Figure 3A). Four trucks with ball bearings contact two rails, to constrain the carriage to move in a nominally linear fashion. Each truck has two loops of balls (Figure 3B), an inner loop and outer loop, with each loop containing 32 balls. Correspondingly, each rail has an inner raceway and an outer raceway (Figure 3C). The two loops interact with different raceways (or grooves) in the rails, one inner and one outer. Whenever the carriage moves back and forth, the balls within the inner/outer loop of a truck rotate in and out of contact with the inner/outer raceway of the rail. At any given instant, about 13 balls per loop contact the given raceway of the rail. Hence, there are about 104 balls (13 balls per loop × 2 loops per truck × 4 trucks) in contact with the two rails.

Each truck was modified to be a "smart truck" equipped with an inductive proximity sensor to detect the phase of the outer loop of balls. Figure 4A shows how each truck (Truck 2 in the figure) was modified with a slot and a tapped access hole, in which resides an inductive proximity sensor. The sensor is used to detect the presence of a metallic or non-metallic ball in its proximity. Instead of using thirty-two chrome steel balls that come with each truck, six of the thirty-two balls were replaced with nylon balls. Figure 4B shows the pattern of twenty-six metal balls with six nylon balls utilized for the outer loop of each truck. Each ball in the pattern, whether metallic or non-metallic, has its own identification number. The pattern was chosen out of many possibilities so that, at any time, visual inspection of the balls through the slot yields a unique pattern of plastic and metal balls for identification. For example, based on Figure 4B, the visible pattern in Figure 4A begins on the left side of the slot with the fourth ball (a nylon ball).

During linear axis motion, the voltage output of the inductive proximity sensor switches with a frequency of 5 kHz between about 0 V (for the presence of nylon) to a nominal voltage when enough metal appears in front of the sensor. A cross-sectional schematic illustration (Figure 4C) reveals that a ball moves to within less than 0.8 mm of the front surface of the inductive proximity sensor, which is sufficient for detectability. By measuring the output voltage as the carriage moves, the phase of the outer loop of balls can be deduced based on the fact that

Vogl, Gregory; Galfond, Brian; Jameson, Jordan.
"BEARING METRICS FOR HEALTH MONITORING OF MACHINE TOOL LINEAR AXES."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

the zero-volt portions of the signal follow a known pattern (Figure 4B).

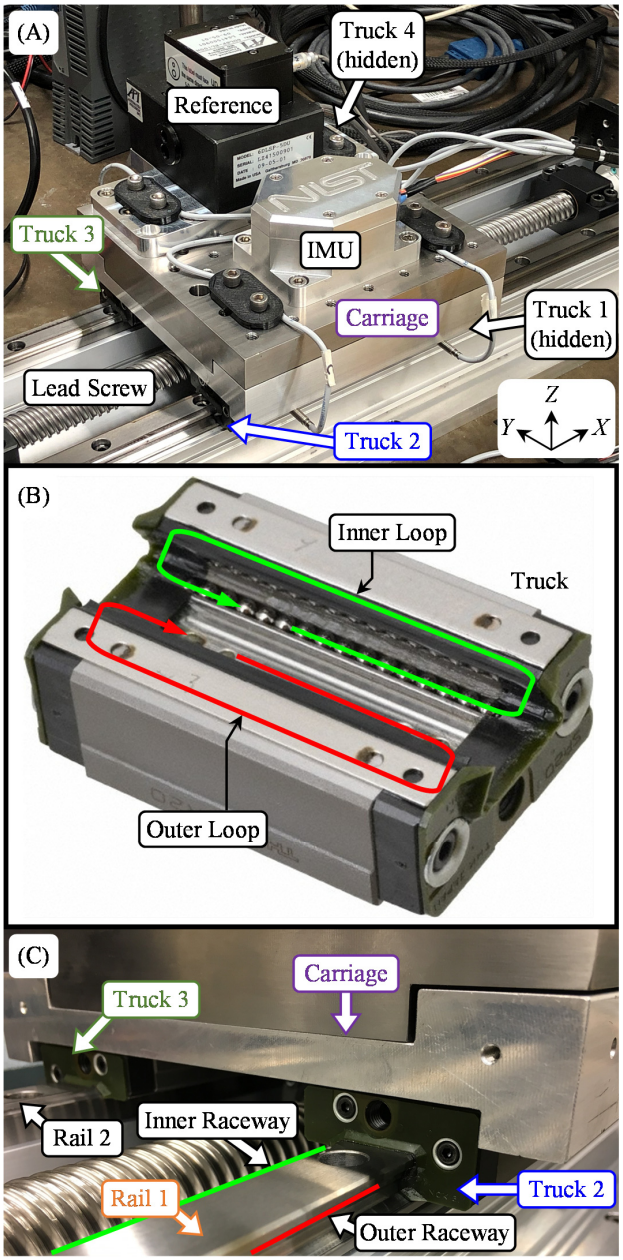


FIGURE 3: (A) LINEAR AXIS WITH IMU, (B) VIEW OF A TRUCK BOTTOM WITH LOOPS OF BALL BEARINGS EXPOSED, WHERE "INNER LOOP" CONTACTS INNER RACEWAY AND "OUTER LOOP" CONTACTS OUTER RACEWAY, AND (C) VIEW OF EXPERIMENTAL SETUP SHOWING TRUCK 2 ATTACHING TO RAIL 1 VIA RACEWAYS.

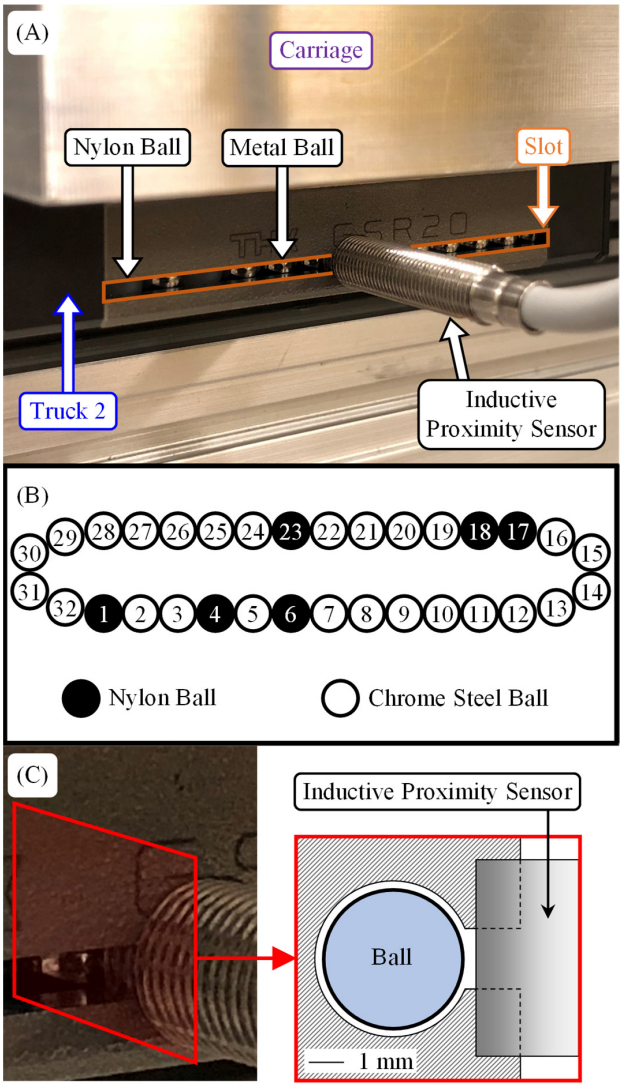


FIGURE 4: (A) VIEW OF A TRUCK WITH EMBEDDED INDUCTIVE PROXIMITY SENSOR, (B) PATTERN OF METALLIC AND NON-METALLIC BALLS USED IN OUTER BALL LOOP OF EACH TRUCK, AND (C) CROSS-SECTIONAL SCHEMATIC ILLUSTRATION OF A TRUCK WITH SENSOR.

3. EXPERIMENTAL PROCEDURE

In this experiment, the metal balls were progressively removed and degraded from the two trucks (Truck 1 and Truck 2) on Rail 1. Figure 5 shows the procedure for damaging a ball one at a time. First, a single ball is removed from the outer loop of a truck (Figure 5A). The tapped hole in the truck allows removal of a ball while leaving the carriage/truck/rail/lead screw system effectively unchanged. This was necessary because if the trucks were removed and then reassembled at each stage of degradation, then the load on each truck would change from one assembly to the next, which would change the error motions as

Vogl, Gregory; Galfond, Brian; Jameson, Jordan.
"BEARING METRICS FOR HEALTH MONITORING OF MACHINE TOOL LINEAR AXES."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

well. Hence, to eliminate the need for disassembly and reassembly of the trucks during the experiment and thus ensure that changes in error motions are due only to damage in outer loop balls, the trucks were modified with a tapped access hole (Figure 4A). Second, the ball is abrasively modified to have a flat with a nominal depth of 30 µm (Figure 5B). Third, as seen in Figure 5C, the nominal diameter and the flat-to-sphere distance of the ball are measured with a micrometer and the difference of the two measurements is an estimate of the flat depth. Specifically, five measurements are performed for both the nominal diameter and the flat-to-sphere distance, and the difference of the respective averages is the estimated flat depth. Finally, the ball is placed back into its position in the outer loop of the truck.

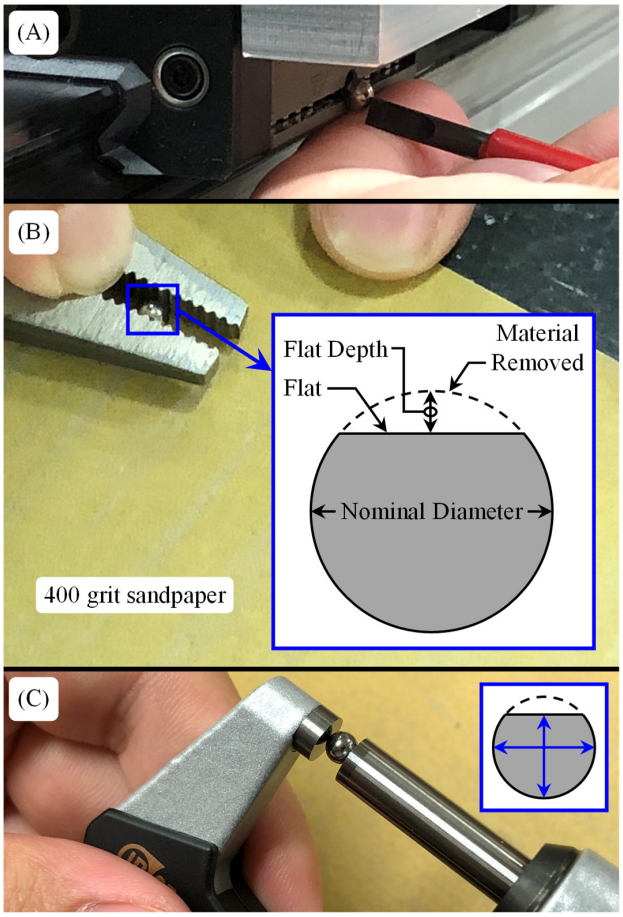


FIGURE 5: PROCEDURE FOR DAMAGING A BALL: (A) REMOVAL OF BALL FROM OUTER LOOP, (B) ABRASIVE REMOVAL OF MATERIAL TO CREATE FLAT, AND (C) ESTIMATION OF FLAT DEPTH WITH HANDHELD MICROMETER MEASUREMENTS.

Over the course of the experiment, each of the metal balls in the outer loop of Truck 2 was damaged in a specific order until each ball had six flats, and then the same process was repeated for Truck 1. Figure 6 shows the order in which the twenty-six balls in the outer loop of a truck were damaged before a "damage cycle" is complete. The seemingly haphazard order was chosen *a priori* as an attempt to simulate random ball damage. Once four balls were damaged and placed back into Truck 2, fifty (50) runs of IMU data were collected. Each run consists of moving the carriage, forward and backward, at three different speeds: 0.02 m/s, 0.1 m/s, and 0.5 m/s. Also, ten (10) bidirectional runs of laser-based reference data were collected statically at 1 mm intervals. Then, another four balls were damaged, and data was collected again for that new "stage" of degradation. Table 1 shows the numbers of the outer loop balls in Truck 2 that were damaged at each of the thirty-nine stages of degradation, based on the damage cycle of Figure 6. Truck 2 was "fully damaged" once six damage cycles were completed; that is, each chrome-steel ball in the outer loop of Truck 2 had six flats (Stage 39 in Table 1). Finally, the same ball-damage process was repeated for Truck 1 (Stage 40 to Stage 55), except that only about 3 flats were induced on each ball in Truck 1. A total of 220 flats were induced on 52 balls (26 balls per truck) in the entire experiment. IMU and laser-based reference data were gathered at each stage of degradation (Stage 0 to Stage 55).

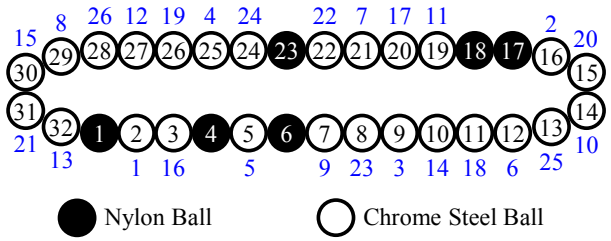


FIGURE 6: ORDER OF OUTER LOOP BALL DAMAGE. THE BLUE NUMBERS REPRESENT THE ORDER IN WHICH THE METAL BALLS ARE DAMAGED. WHENEVER A CYCLE IS COMPLETE (ALL TWENTY-SIX METAL BALLS HAVE SAME NUMBER OF FLATS), A NEW CYCLE MAY BEGIN.

Figure 7 shows histograms of the flat depths for all 220 flats induced on the metal balls in Truck 2 and Truck 1. Except for one outlier (not shown in Figure 7A), the flat depth ranges between 25 µm and 35 µm with an average of approximately 30 µm. Specifically, the flat depth in Figure 7 is the difference between the average nominal diameter, resulting from five measurements, and the average flat-to-sphere distance, also from five measurements (see Figure 5C). The five nominal diameter measurements had an average standard deviation of 0.44 µm, while the flat-to-sphere measurements had an average standard deviation of 0.92 µm. The greater standard deviation for the flat-to-sphere measurements is due in large part to the non-flatness of the flat (the "flat" is only nominally flat) combined with human operation of the micrometer.

Vogl, Gregory; Galfond, Brian; Jameson, Jordan.
"BEARING METRICS FOR HEALTH MONITORING OF MACHINE TOOL LINEAR AXES."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

TABLE 1: NUMBER OF BALLS IN OUTER LOOP OF TRUCK 2 INDUCED WITH FLATS AT EACH STAGE OF DEGRADATION.

Stage	Damaged Balls		Stage	Damaged Balls
0	None		20	13, 28, 2, 16
1	2, 16, 9, 25		21	9, 25, 5, 12
2	5, 12, 21, 29		22	21, 29, 7, 14
3	7, 14, 19, 27		23	19, 27, 32, 10
4	32, 10, 30, 3		24	30, 3, 20, 11
5	20, 11, 26, 15		25	26, 15, 31, 22
6	31, 22, 8, 24		26	8, 24, 13, 28
7	13, 28, 2, 16		27	2, 16, 9, 25
8	9, 25, 5, 12		28	5, 12, 21, 29
9	21, 29, 7, 14		29	7, 14, 19, 27
10	19, 27, 32, 10		30	32, 10, 30, 3
11	30, 3, 20, 11		31	20, 11, 26, 15
12	26, 15, 31, 22		32	31, 22, 8, 24
13	8, 24, 13, 28		33	13, 28, 2, 16
14	2, 16, 9, 25		34	9, 25, 5, 12
15	5, 12, 21, 29		35	21, 29, 7, 14
16	7, 14, 19, 27		36	19, 27, 32, 10
17	32, 10, 30, 3		37	30, 3, 20, 11
18	20, 11, 26, 15		38	26, 15, 31, 22
19	31, 22, 8, 24		39	8, 24, 13, 28

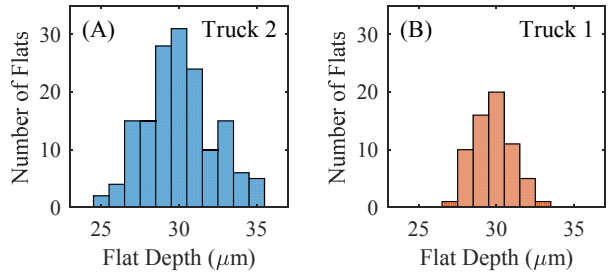


FIGURE 7: HISTOGRAM OF FLAT DEPTH FOR (A) TRUCK 2 AND (B) TRUCK 1.

The nominal flat depth of 30 µm was chosen to be large enough to influence the geometric error motions of the linear axis, without being too large as to be unrealistic of possible damage within a machine tool linear axis. The volume, $\boldsymbol{v}$, of material removed by a flat is defined as

$$v = \frac{\pi}{3} d^2 (3r - d) \tag{1}$$

and the curved area, a, of the ball marred with the flat is

$$a = 2\pi r d \tag{2}$$

where d is the flat depth and r is the nominal ball radius. The metal balls were made of chromium steel with a nominal diameter of 3.965 mm. For a nominal flat depth, d, of 30 µm (see Figure 7), Eq. (1) shows that the nominal volume of metal material removed by a flat is about 0.0056 mm^3, which is a relatively small volume. Also, Eq. (2) shows that the nominal area of metal marred by the flat is about 0.37 mm^2, which is a relatively small area. However, after 220 flats have been induced, the cumulative material removal is approximately 1.23 mm^3 and the cumulative marred area is about 0.82 cm^2. These flats interact with the rails to influence the geometric error motions.

Theoretically, the surface roughness of the truck balls (not material volume) and the total damaged surface area are metrics that correlate with the geometric error motions, specifically the straightness and angular error motions. The average surface roughness, R_a, across all 208 metal balls in the trucks (4 trucks × 2 loops per truck × 26 metal balls per loop) can be estimated via the material removal as

$$R_\mathrm{a} \approx \bar{R}_\mathrm{a} + \frac{V}{A_\mathrm{T}} \tag{3}$$

where $\bar{R}_\mathrm{a}$ is the nominal average surface roughness [11] excluding the flats, V is the total cumulative volume of material removed by all flats at any moment, and A_T is the total cumulative surface area of all metal balls in the trucks (approximately equal to 103 cm^2). The balls were specified as Grade 25, which has a maximum allowable average surface roughness of 0.0508 µm, so we assume that $\bar{R}_\mathrm{a}$ = 0.0508 µm. Consequently, after 220 flats have been induced for a total material removal of 1.23 mm^3, Eq. (3) reveals that the average surface roughness across all metal truck balls is then approximately 0.17 µm. Similarly, the relative damaged surface area, A_rel, is defined as

$$A_\mathrm{rel} = \frac{A}{A_\mathrm{T}} \tag{4}$$

where A is the total cumulative damaged ball area due to all flats at any moment. After 220 flats have been induced, Eq. (4) reveals that approximately 0.80% of the metal ball surface area has been damaged by the numerous flats. Thus, the 220 flats damage almost 1% of the total surface area, in addition to tripling the average surface roughness, of the metal balls.

This experiment and subsequent analysis explores the ability for such a significant change in surface roughness to be detected by an IMU-based methodology, based on changes detected in the geometric error motions of the linear axis. Figure 8 shows the surface roughness and relative damaged surface area calculated according to Eq. (3) and Eq. (4), respectively, for all stages of degradation. Because the depth of each flat is nominally 30 µm (Figure 7), the surface roughness and relative damaged surface area change relatively linearly with degradation stage.

Note that the nylon balls are not included in the surface roughness equation, Eq. (3), because the nylon balls bear insignificant loads. The nylon balls had a nominal diameter of 3.969 mm, which is 4 µm greater than the nominal diameter of 3.965 mm for the chromium steel balls. Therefore, the nylon balls will bear some nominal load. However, because nylon has such a relatively low modulus of elasticity compared to that for chromium steel, the nylon balls deflect easily under very low

Vogl, Gregory; Galfond, Brian; Jameson, Jordan.
"BEARING METRICS FOR HEALTH MONITORING OF MACHINE TOOL LINEAR AXES."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

loads. For example, the relatively small force due to the ratchet of the micrometer deflected the nylon balls about 72 µm. The low stiffness of the nylon balls means that the nylon balls act as metal-ball spacers without otherwise significantly influencing the geometric error motions.

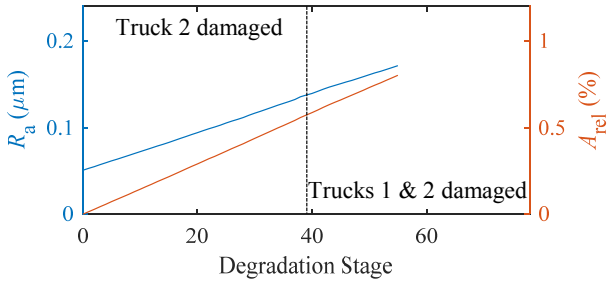


FIGURE 8: AVERAGE SURFACE ROUGHNESS AND RELATIVE DAMAGED SURFACE AREA OF ALL METAL BALLS WITHIN TRUCKS VERSUS DEGRADATION STAGE.

4. FEATURES FOR DATA ANALYSIS

The IMU and laser-based reference data that were gathered at each stage of degradation (Stage 0 to Stage 55) can now be used for analysis. The fifty (50) runs of time-sampled IMU data were processed to yield the error motions as functions of travel position (X values) [8, 9], while the ten (10) runs of laser-based reference data were already functions of position, being collected at nominal positions with an interval of 1 mm. Based on the rate gyroscope and accelerometer bandwidths and axis speed, the available spatial frequency range is between 0 and 2 cycles/mm (2000 m^{-1}). Hence, for each of the stages of degradation, there are fifty (50) IMU-based data runs and ten (10) reference-based data runs available for analysis.

The statistical time-domain features used in the analysis of this experiment are described in detail in Ref. [12]: peak value (PV), root-mean-square (RMS), standard deviation (σ), skewness (γ_1), kurtosis (γ_2), crest factor (CF), shape factor (SF), impulse factor (IF), and clearance factor (CLF). The feature formulas are given in Table 2, where $\bar{y}$ denotes the average of a signal (y) and N denotes the number of data points in the signal. For the current implementation, each 'signal' is a geometric error motion as a function of position, or some transformation of the error motion, e.g., via filtering. The features were calculated for every instance of error motion data generated at each stage of degradation.

The features in Table 2 were applied to mean-subtracted, band-pass filtered signals. First, the error motions were bandpass filtered between spatial frequencies of 200 m^{-1} and 10^4 m^{-1}, which have cutoff wavelengths of 5 mm and 0.1 mm, respectively. The cutoff wavelength of 0.1 mm is the smallest wavelength achievable, corresponding to the bandwidth of the rate gyroscopes (200 Hz) and the axis speed used for that frequency range (0.02 m/s). On the other hand, the cutoff wavelength of 5 mm is sufficient to capture any impacts from bearing faults, since bearing fault signatures are assumed to occur with wavelengths on the order of about 0.7 mm (the nominal diameter of the flats) and less than about 12.4 mm (circumference of a single ball). Filtering was performed with a first-order, zero-phase digital Butterworth filter [13]. Second, to eliminate any portion of data that relates to the shape of the rail raceways, the mean of the filtered signals was removed. This means that for the fifty (50) runs of IMU-based data for each of the six degrees of freedom, the means of all 50 filtered signals were subtracted from each individual signal for feature analysis. Similarly, the filtering and mean-removal processes were applied to the ten (10) laser-based reference data for each stage of degradation.

TABLE 2: STATISTICAL FEATURES.

Feature Name	Formula
Peak Value	$PV(y) = \frac{1}{2}\left(\max(y) - \min(y)\right)$
Root-Mean-Square	$RMS(y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} y_i^2}$
Standard Deviation	$\sigma(y) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$
Skewness	$\gamma_1(y) = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^3}{\sigma^3}$
Kurtosis	$\gamma_2(y) = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^4}{\sigma^4}$
Crest Factor	$CF(y) = \frac{PV(y)}{rms(y)}$
Shape Factor	$SF(y) = \frac{rms(y)}{\frac{1}{N}\sum_{i=1}^{N}\lvert y_i\rvert}$
Impulse Factor	$IF(y) = \frac{PV(y)}{\frac{1}{N}\sum_{i=1}^{N}\lvert y_i\rvert}$
Clearance Factor	$CLF(y) = \frac{PV(y)}{\left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\lvert y_i\rvert}\right)^2}$

5. DATA ANALYSIS AND RESULTS

An example of a feature applied to the IMU-based data is Figure 9, which shows the standard deviation of each filtered error motion. There are 55 degradation stages, and there are fifty (50) runs at each stage, leading to fifty values of standard deviation at each stage. Specifically, the standard deviation in Figure 9 is the standard deviation of a bandpass filtered version of an error motion (see Section 4), denoted by a tilde over each error motion, e.g., $\tilde{E}_{XX}$ and $\tilde{E}_{AX}$. At each stage, the blue box represents the middle 50 percent of the 50 values, and the black whiskers extend to the largest value that falls below $q_3 + 1.5(q_3 - q_2)$, or the smallest value that falls above $q_2 - 1.5(q_3 - q_2)$, where q_2 and q_3 are the 25th and 75th percentile respectively. Values that fall outside this range are classified as outliers, which are plotted as blue crosses in Figure 9. Also, the red line within each blue box is the median value for all 50 runs for that degradation stage.

Vogl, Gregory; Galfond, Brian; Jameson, Jordan.
"BEARING METRICS FOR HEALTH MONITORING OF MACHINE TOOL LINEAR AXES."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

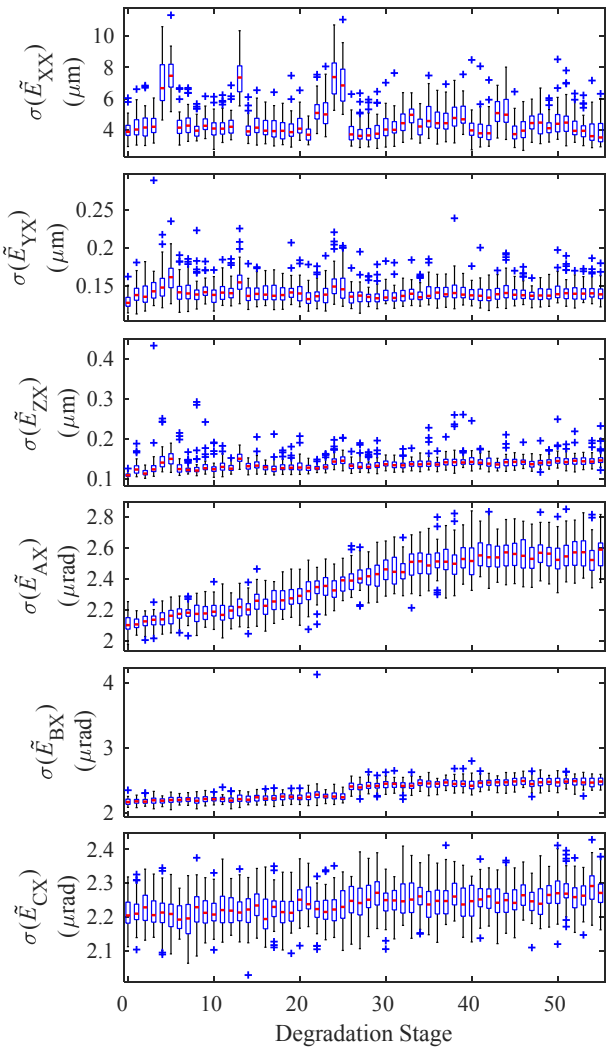


FIGURE 9: BOXPLOTS OF STANDARD DEVIATION OF FILTERED ERROR MOTION VERSUS DEGRADATION STAGE.

Some trends of the metric, $\sigma(\tilde{E}_{AX})$, with degradation stage are evident in Figure 9, while others are more unclear. For example, one clear trend is that $\sigma(\tilde{E}_{AX})$ increases fairly linearly from a value of about 2.1 µrad at Stage 0 to about 2.5 µrad at Stage 40, which is a 0.4 µrad increase in the metric. Then, the metric essentially stays constant after Stage 40. This difference in trends (linear versus constant metric) is related to which balls were damaged at each stage: metal balls of Truck 2 were only damaged until Stage 40, at which point new flats were induced only on the metal balls of Truck 1. Therefore, the difference in trends is due to the dependence of $\sigma(\tilde{E}_{AX})$ on the balls of the trucks, accounting for the linear trend before Stage 40 (due to damage within Truck 2) and the constant trend after Stage 40 (due to damage within Truck 1). Perhaps the physical mechanism that leads to the trend difference is the preload on each truck, with the preload being much greater on the balls of Truck 2 compared to Truck 1. A greater preload within Truck 2 could lead to a greater dependence of E_{AX} on changes to the balls within Truck 2.

One way to assess trends is to monitor a characteristic of a statistical population of the feature with degradation stage. Each of the boxplots in Figure 9 represents a statistical population at a given stage of degradation. As such, the statistical characteristics (mean, median, etc.) of each population can be monitored from stage to stage. Because median trends are visible in Figure 9, and the data has relatively large (but few) outliers, the median value is chosen instead of the mean for monitoring purposes. For the standard deviation feature, we can approximate the contribution of degradation to the median value. Based on the assumption that the additional signal (from degradation) and baseline signal (without degradation) are uncorrelated, this relationship can be represented as

$$\sigma_i^2 = \sigma_0^2 + \sigma_{d,i}^2 \tag{5}$$

where σ_0 is the standard deviation without degradation, and $\sigma_{d,i}$ is the contribution from degradation at the i^{th} stage of degradation. To ensure that the contribution from degradation is positive and begins at zero at Stage 0, the median contribution is approximated as

$$\hat{\sigma}_{d,i} = \sqrt{\hat{\sigma}_i^2 - \min(\hat{\sigma}_i)^2} - \hat{\sigma}_0 \tag{6}$$

where a "hat" denotes the median value, and the minimum function is for all median values. Equation (6) can be applied to the IMU-based data as well as the laser-based reference data.

Figure 10 shows the approximate contribution of degradation to the median standard deviation, $\hat{\sigma}_{d,i}$, for the IMU- and reference-based data for one DOF (roll angle). Even though the reference data has a lower noise than the IMU data (Figure 10A), the trend of the approximate contribution from degradation is very similar and within 0.5 µrad (Figure 10B). There appears to be a scaling difference between the IMU- and reference-based results for $\hat{\sigma}_d(\tilde{E}_{AX})$, which could be due to the differences in data collection (carriage dynamics, sensor bandwidths, thermal states, etc.). For the other five DOF, the differences are much larger than a scaling factor. Figure 11 compares the IMU- and reference-based approximate contribution of degradation for all six DOF. The medians trend at significantly different rates (for $\tilde{E}_{YX}$ and $\tilde{E}_{BX}$) or even trend in opposite directions (for $\tilde{E}_{ZX}$ and $\tilde{E}_{CX}$). A question then arises: How do we assess trends of features with degradation stage?

One way to assess the trends of features is to evaluate the correlations with degradation stage. The Pearson correlation coefficient, ρ_P, measures the linearity of a trend, while the Spearman correlation coefficient, ρ_S, measures the monotonicity of a trend [14]. Because both correlation coefficients have merit,

Vogl, Gregory; Galfond, Brian; Jameson, Jordan.
"BEARING METRICS FOR HEALTH MONITORING OF MACHINE TOOL LINEAR AXES."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

the geometric mean of their magnitudes, $\sqrt{|\rho_P||\rho_S|}$, can be used to quantitatively assess a trend. The closer the geometric mean is to the maximum possible value of one, the greater the correlation of a trend with degradation stage.

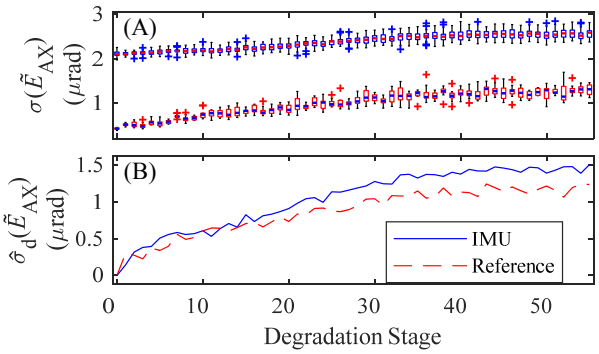


FIGURE 10: (A) BOXPLOTS OF STANDARD DEVIATION OF FILTERED ERROR MOTION BASED ON IMU DATA (BLUE BOXES AND CROSSES) AND REFERENCE DATA (RED BOXES AND CROSSES), AND (B) RELATIVE MEDIAN VALUES.

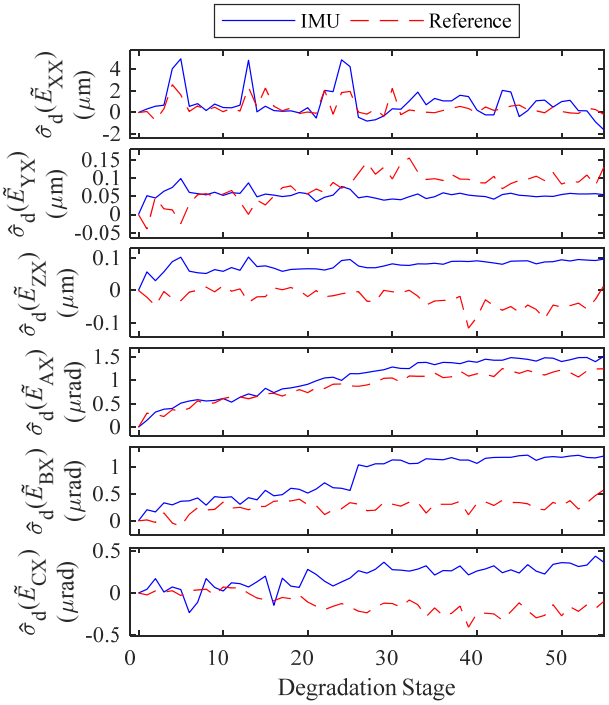


FIGURE 11: (A) RELATIVE MEDIAN VALUES FOR IMU- AND REFERENCE-BASED RESULTS.

The metric, $\sqrt{|\rho_P||\rho_S|}$, can be calculated for any statistical characteristic of any feature. Figure 12 shows the geometric mean of the Pearson and Spearman coefficients for the median of each of the nine statistical features (Table 2). Each of the six degrees of freedom has a different value of $\sqrt{|\rho_P||\rho_S|}$ for each of the nine features. Visual inspection shows that there can be significant variety among the metric values for the degrees of freedom, leading to a "skyscraper" effect in Figure 12. Nonetheless, some features trend with the degradation of the outer loop balls, since many of the geometric means in Figure 12 have magnitudes above 0.5.

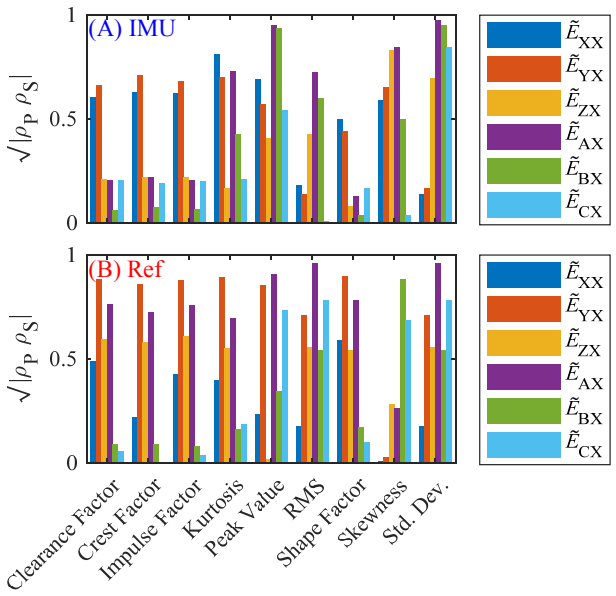


FIGURE 12: GEOMETRIC MEAN OF PEARSON AND SPEARMAN CORRELATION COEFFICIENTS FOR MEDIAN OF EACH STATISTICAL FEATURE BASED ON DATA FROM (A) IMU AND (B) REFERENCE.

In order to quantitatively assess the metrics in the aggregate for a feature, the arithmetic mean, $\bar{\rho}$, is defined as

$$\bar{\rho} = \text{mean}\left(\sqrt{|\rho_P||\rho_S|}\right) \text{ for six values for each feature} \quad (7)$$

Figure 13 compares the arithmetic mean of the IMU- band reference-based medians for the nine statistical features (Table 2). The IMU-based values of $\bar{\rho}$ range from about 0.3 (shape factor feature) to about 0.7 (peak value and standard deviation feature), while reference-based values of $\bar{\rho}$ range from about 0.5 (skewness feature) to about 0.7 (RMS and standard deviation features). Therefore, the standard deviation feature has, on average, the largest value of $\bar{\rho}$ and is therefore chosen as the feature to be investigated for monitoring purposes.

Not only can the median value of standard deviations be used for monitoring trends (Figure 11) in a physical sense of micrometers or microradians, but the Wilcoxon rank sum test can be used to determine whenever the median value of one population (for Stage 1 and beyond) has deviated significantly from that for the initial population (for Stage 0) [14]. If the Wilcoxon rank sum test probability, P, is below 0.05 (5%), the

Vogl, Gregory; Galfond, Brian; Jameson, Jordan.
"BEARING METRICS FOR HEALTH MONITORING OF MACHINE TOOL LINEAR AXES."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

distribution for that stage is statistically different from the initial stage. In that case, the hypothesis, H, equals 1, because the hypothesis of a statistically significant change due to degradation is true. On the other hand, whenever the probability, P, is not below 0.05 (5%), the hypothesis is false ($H = 0$), since the median value of the given distribution has not changed much from the initial value for Stage 0; degradation is unclear.

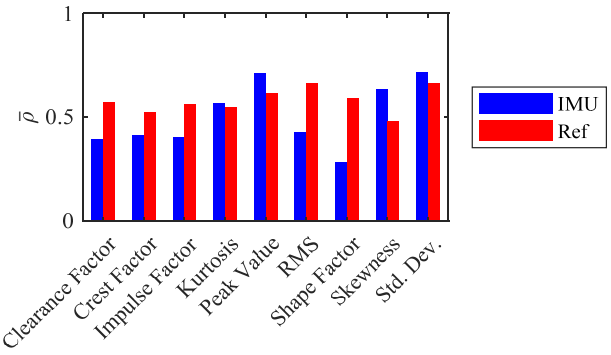


FIGURE 13: ARITHMETIC MEAN OF GEOMETRIC MEANS FOR MEDIAN OF EACH STATISTICAL FEATURE OF IMU-BASED AND REFERENCE-BASED RESULTS.

Figure 14 shows the probability, P, and hypothesis, H, values for $\sigma(\tilde{E}_{YX})$. The IMU detects a change at Stage 1, which is much earlier than when the reference detects a change at Stage 16, despite the reference yielding a much greater correlation ($\sqrt{|\rho_P||\rho_S|} = 0.7$) than the IMU ($\sqrt{|\rho_P||\rho_S|} = 0.2$) for standard deviation of $\tilde{E}_{YX}$ in Figure 12. To smooth abrupt changes of P for condition monitoring purposes, filtered versions of P and H, denoted as P_F and H_F, are also shown in Figure 14. The P curve is filtered with a first-order Savitzy-Golay filter of length 11 to yield P_F, which then results in H_F via the 5% threshold. In Figure 14A, the H_F value switches at Stage 6 because the Savitzy-Golay filter of length 11 (= 5 + 1 + 5) smooths out the abrupt shift of P at Stage 1 to last until Stage 6 (= 1 + 5).

Figure 15 shows the P_F curves for standard deviation for all degrees of freedom. The IMU-based results cross the 5% threshold only once, while the reference-based results sometimes cross back and forth across the 5% threshold, despite increasing degradation. Furthermore, the IMU-based results cross the 5% threshold at earlier stages than the reference-based results, showing how the IMU is perhaps more sensitive and better suited for monitoring purposes. The only degree of freedom that does not cross the 5% threshold is $\tilde{E}_{XX}$, because the positioning error motion is not significantly affected by truck ball degradation.

Figure 15 can be combined with Figure 11 to determine the minimum significant changes of standard deviation for five DOF (positioning error excluded). The IMU-based values in Figure 11 at the threshold crossings in Figure 15 are either about 0.05 µm (for translational error motions) or 0.5 µrad (for angular error motions). Therefore, 0.05 µm and 0.5 µrad are the minimum statistically-significant changes of standard deviation due to degradation for the IMU-based results. In fact, these two minima are essentially the same, because the characteristic distance between trucks is roughly 0.1 m, which means that a change of 0.05 µm would create an angular change of about 0.5 µrad (= 0.05 µm/0.1 m). Finally, the crossing locations in Figure 15A range from Stage 6 to Stage 28, which relate to 0.09% or 0.4% of damage to the total metal ball surface area. Consequently, as little as 0.09% of total metal ball area damage can yield a statistically significant change of standard deviation.

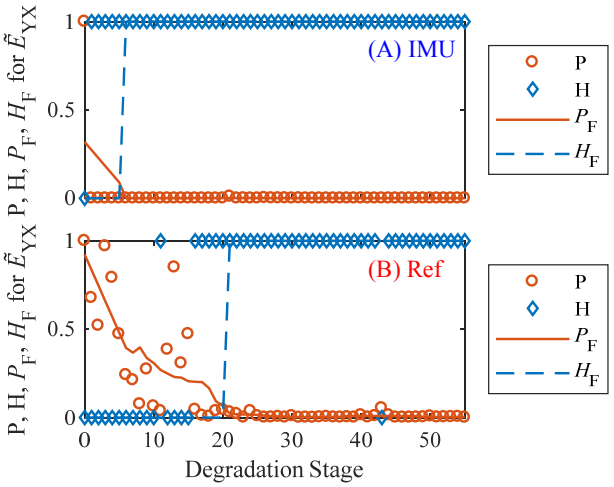


FIGURE 14: WILCOXON RANK SUM TEST, AND ITS FILTERED FORM, FOR $\sigma(\tilde{E}_{YX})$ BASED ON DATA FROM (A) IMU AND (B) REFERENCE.

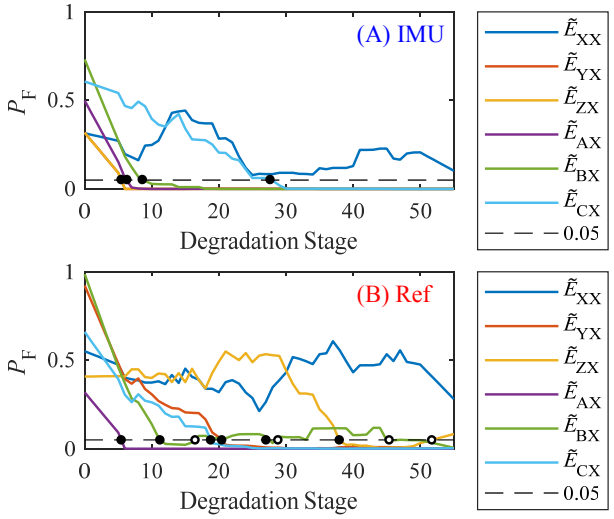


FIGURE 15: FILTERED VERSION OF WILCOXON RANK SUM TEST VERSUS DEGRADATION STAGE, BASED ON DATA FROM (A) IMU AND (B) REFERENCE. DOWNWARD/UPWARD CROSSINGS WITH THE (DASHED) FIVE PERCENT LINE ARE DENOTED WITH BLACK/WHITE FILLED CIRCLES.

Vogl, Gregory; Galfond, Brian; Jameson, Jordan.
"BEARING METRICS FOR HEALTH MONITORING OF MACHINE TOOL LINEAR AXES."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

6. CONCLUSIONS

An inertial measurement unit was studied for its ability to measure changes in geometric error motions due to induced faults on the recirculating ball bearings of two carriage trucks within a linear axis. Each truck was modified with an access hole that allowed the metal balls in the outer loop of the truck to be progressively removed and degraded. In this experiment, each of the metal balls in the outer loop of two trucks were damaged in a specific order until each ball had six flats, each with a nominal depth of 30 µm. For each stage of degradation, four new flats were induced and then fifty (50) runs of IMU data were collected in addition to ten (10) runs of laser-based reference data. After 220 flats had been induced, approximately 0.80% of the metal ball surface area was damaged by the numerous flats, and the average surface roughness was changed by about 0.11 µm.

For each stage of degradation, the fifty runs of IMU data and ten runs of reference data were analyzed with the nine features (standard deviation, kurtosis, etc.) used for rotating machinery systems. Trends in the median value of the statistical populations were determined via the Pearson and Spearman correlation coefficients. The standard deviation feature had one of the greatest correlations with degradation among both the IMU and reference data, so the standard deviation feature was used for the rest of the analysis. Subsequently, the Wilcoxon rank sum test was used to reveal an ability of the standard deviation feature to detect statistically significant changes as small as 0.05 µm or 0.5 µrad, corresponding to a total damaged surface area of truck balls of less than 0.1 percent.

Therefore, results showed that the IMU-based monitoring system has promise for online, data-rich, integrated diagnostics and prognostics of linear axes system health. Feature changes due to increasing degradation were identified in the median of standard deviation and other features. Pearson and Spearman correlation analysis provided a high-level view of these trends and insight into how to select a feature for monitoring. On the other hand, the Wilcoxon rank sum test provided a low-level view of how to determine statistical lower bounds, for thresholding purposes, within future online monitoring systems.

Future work includes analysis of filtered components with spatial frequencies that are different than those used in this study, as well as the incorporation of the inductive proximity sensor data. The “smart truck” concept, used to detect the phase of the outer loop of balls, could be used to study the influence of the balls and their damage on the geometric error motions.

ACKNOWLEDGEMENTS

The authors thank Casey Shatzley and the Fabrication Technology Office (National Institute of Standards and Technology) for modifying the trucks used in the experiments.

NIST DISCLAIMER

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States. Approved for public release; distribution is unlimited.

REFERENCES

[1] Altintas Y, Verl A, Brecher C, Uriarte L, Pritschow G (2011) Machine Tool Feed Drives. *CIRP Annals-Manufacturing Technology* 60(2):779-796.

[2] Li Y, Wang X, Lin J, Shi S (2014) A Wavelet Bicoherence-Based Quadratic Nonlinearity Feature for Translational Axis Condition Monitoring. *Sensors (Basel)* 14(2):2071-2088.

[3] Zhou Y, Mei X, Zhang Y, Jiang G, Sun N (2009) Current-Based Feed Axis Condition Monitoring and Fault Diagnosis. 4th IEEE Conference on Industrial Electronics and Applications, ICIEA 2009, 1191-1195.

[4] Uhlmann E, Geisert C, Hohwieler E (2008) Monitoring of Slowly Progressing Deterioration of Computer Numerical Control Machine Axes. *Proceedings of the Institution of Mechanical Engineers Part B-Journal of Engineering Manufacture* 222(10):1213-1219.

[5] International Organization for Standardization (2012) ISO 230-1 - Test Code for Machine Tools − Part 1: Geometric Accuracy of Machines Operating under No-Load or Quasi-Static Conditions.

[6] Khan AW, Chen W (2008) Calibration of CNC Milling Machine by Direct Method. 2008 International Conference on Optical Instruments and Technology: Optoelectronic Measurement Technology and Applications, 7160:716010.

[7] Teti R, Jemielniak K, O'Donnell G, Dornfeld D (2010) Advanced Monitoring of Machining Operations. *CIRP Annals-Manufacturing Technology* 59(2):717-739.

[8] Vogl GW, Donmez MA, Archenti A (2016) Diagnostics for Geometric Performance of Machine Tool Linear Axes. *CIRP Annals-Manufacturing Technology* 65(1):377-380.

[9] Vogl GW, Donmez MA, Archenti A, Weiss BA (2016) Inertial Measurement Unit for On-Machine Diagnostics of Machine Tool Linear Axes. Annual Conference of the Prognostics and Health Management Society 2016, 7.

[10] Vogl GW, Sharp ME (2017) Diagnostics of Machine Tool Linear Axes Via Separation of Geometric Error Sources. Annual Conference of the Prognostics and Health Management Society,

[11] Black JT, Kohser RA (2012) *DeGarmo's Materials and Processes in Manufacturing*. 11th ed. John Wiley & Sons, Inc., USA.

[12] Jameson J, Vogl GW (2018) Comparative Analysis of Bearing Health Monitoring Methods for Machine Tool Linear Axes. Society for Machinery Failure Prevention Technology 2018 (MFPT 2018), 1-16.

[13] Oppenheim AV, Schafer RW, Buck JR (1999) *Discrete-Time Signal Processing*. 2nd ed. Prentice Hall, Upper Saddle River, NJ.

[14] Gibbons JD, Chakraborti S (2003) *Nonparametric Statistical Inference*. 4th ed. Marcel Dekker Inc., New York.

Vogl, Gregory; Galfond, Brian; Jameson, Jordan.
”BEARING METRICS FOR HEALTH MONITORING OF MACHINE TOOL LINEAR AXES.”
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

2nd International Conference on Natural Hazards & Infrastructure
23-26 June, 2019, Chania, Greece

Selecting Building Characteristics to Predict Seismic Retrofit Costs of a Building Portfolio

J.F. Fung[1], S. Sattar, D.T. Butry, S. McCabe
National Institute of Standards and Technology

ABSTRACT

An accurate yet simple estimate of the retrofit cost plays an important role in the decision-making process of retrofitting existing buildings. Fung et al. (2018b) developed a predictive model to estimate seismic retrofit costs as a function of building characteristics such as building area, building age, and building model type. However, in practice, a decision maker may not have access to the full set of building characteristics required for estimating the retrofit cost, especially when dealing with a portfolio of buildings. Certain characteristics (e.g., building area) might be more readily available or easier to obtain than others (e.g., building type). This paper considers the tradeoff, in terms of prediction error, from not using all of the building characteristics necessary for prediction of retrofit cost. The results show that excluding certain characteristics from prediction, such as building type, lead to negligible increases in the prediction error. The paper also finds the minimal set of building characteristics needed to approximate the accuracy of the model that uses the full set of building characteristics. Findings of this study will help decisions makers to estimate retrofit costs without having to spend additional time and money to collect the full set of data on the building portfolio.

Keywords: seismic retrofit, retrofit cost prediction, GLM, regularization, variable selection

INTRODUCTION

One option for obtaining seismic retrofit cost estimates for a building is to hire an engineering consulting professional. However, even for a single building, the cost estimate requires detailed structural information, including the existing seismic detailing and material properties. Obtaining this information requires that the engineering professional examine the building on site. The process becomes time consuming and expensive as the number of buildings increases.

An alternative is to estimate retrofit costs for a building based on *historical* retrofit costs of buildings that have been retrofitted. Fung et al. (2018b) develop a predictive model to estimate seismic retrofit costs as a function of eight predictors, shown in Table 1. The advantage of this approach over hiring a professional is that the data required for prediction is generally easier to obtain and does not require on site inspections. Thus, retrofit cost predictions can be generated quickly and cheaply. On the other hand, such cost predictions may not be as accurate as those from an engineering professional.

In practice, some of the building characteristics presented in Table 1 may not be available for some or all buildings in a portfolio. Fung et al. (2018c), for instance, use the predictive modeling approach to obtain seismic retrofit cost estimates for a portfolio of buildings that does not include building age, height, or type.

[1] Corresponding Author: J.F. Fung, *National Institute of Standards and Technology, juan.fung@nist.gov*

Fung, Juan; Sattar, Siamak; Butry, David; McCabe, Steven.
"Selecting Building Characteristics to Predict Seismic Retrofit Costs of a Building Portfolio."
Paper presented at 2nd International Conference on Natural Hazards & Infrastructure, Chania, Greece. June 23, 2019 - June 26, 2019.

Table 1. Target outcome, *Y*, and set of predictors, *X*, based on the retrofit cost model in Fung et al. (2018b).

Variable	Definition	Scale/Values
Y	Retrofit cost	Dollars per square foot
s	Seismicity	{Low, Medium, High, Very High}
p	Performance objective	{Life Safety, Damage Control, Immediate Occupancy}
b	Building group	{1, …, 8}[a]
Area	Total building area	Square feet
Age	Building age	Years
Height	Building height	Total above and below ground stories
Occup	Occupancy during retrofit	{Vacant; In-place; Temporarily relocated within building}
Historic	Building historic status	Is building deemed historic? (Yes or No)

[a] Building model types are categorized into eight building groups, as shown in Table 2.

Note that the performance objective, p, i.e., the target building performance after the retrofit, and the occupancy during retrofit, Occup, are the only predictors that correspond to retrofit *actions*. The other predictors are building characteristics. The performance objective categories represented in Table 1 are defined in FEMA (1994) as follows: Life Safety (LS) "allows for unrepairable damage as long as life is not jeopardized and egress routes are not blocked;" Damage Control (DC) "protects some feature or function of the building beyond life-safety, such as protecting building contents or preventing the release of toxic material;" and Immediate Occupancy (IO) "allows only minimal post-earthquake damage and disruption, with some nonstructural repairs and cleanup done while the building remains occupied and safe."

Two Motivating Questions

This paper explores the performance of the predictive model developed in Fung et al. (2018b) when some of the building characteristics are not available. The paper addresses two motivating questions:

1. What is the *minimal model*, that is, minimal in the number of predictors, for obtaining performance comparable to the *benchmark model* that includes all of the predictors?
2. What is the effect on performance from using a model that deliberately omits building age, height, and type (the *practical model*)?

To answer these questions, the paper compares three candidate models, the *benchmark model*, the *minimal model*, and the *practical model* by their performance in terms of *prediction error*. The next section describes the predictive modeling approach, and the following section presents the results.

METHODOLOGY

Fung et al. (2017, 2018a, 2018b) develop a predictive model for structural seismic retrofit costs, Y, as a function of the predictors shown in Table 1, $Y = f(X)$. The goal is to use historical data X to estimate a function $\hat{f}$ such that $\hat{Y} = \hat{f}(X_{new})$ is a reasonable prediction of retrofit costs for a new building with characteristics X_{new}.

This paper uses a Generalized Linear Model (GLM) to estimate f, as in Fung et al. (2018b).[2] Moran et al. (2007) discuss the use of GLMs for cost prediction. One key advantage of using GLMs, rather than standard linear regression models, for cost prediction is that results are easily interpretable in dollar terms (Fung et al., 2018b).

[2] The paper uses a GLM with gamma-distributed outcome and a log link. The "Linear" part of a GLM means f is a function of a linear combination of the predictors, $X\beta$, where β is a vector of coefficients; thus, $f(X) = e^{X\beta}$. For details, see Fung et al. (2018b).

Given an estimator $\hat{f}$, model performance is estimated using *prediction error*. This paper uses Root Mean Square Error (RMSE), given in Equation (1), as the measure of the prediction error,

$$RMSE(\hat{f}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{f}(X_i) - Y_i)^2}\,, \tag{1}$$

where Y_i is the actual retrofit cost for building *i* and $\hat{f}(X_i)$ is the predicted retrofit cost for building *i* and *m* is the number of observations. Note that both the actual cost and the predicted cost are in the scale of dollars per square foot and thus RMSE is easily interpretable on the same scale.

Prediction error is an important criterion for two distinct model development steps: *model selection* (choosing the best model out of a set of candidate models) and *model evaluation* (estimating a model's expected performance on new data, or "out-of-sample performance"). This paper uses *nested K-fold cross-validation* in order to perform both model selection and model evaluation (Fung et al., 2018b). The approach prevents the data used for selecting a model from contaminating the data used for model evaluation. This is especially useful when the number of observations used to fit, or *train*, the models is small.[3]

In order to address the two motivating questions, this paper considers three models:

1. The *benchmark model*, $\hat{f}_{benchmark}$, is the model that includes all of the predictors in Table 1.
2. The *minimal model*, $\hat{f}_{minimal}$, the model that includes the minimal number of predictors for obtaining performance close to that of the benchmark model.
3. The *practical model*, $\hat{f}_{practical}$, is the model that deliberately omits predictors that may not be easily available (in this case, building age, building height, and building type).

The three models are compared based on their out-of-sample prediction error. Thus, for instance, any increase in prediction error from omitting building age, height, and type can be interpreted as the cost of not collecting information on these predictors.

Regularization

In general, increasing the number of predictors in a regression model tends to improve performance on the *training data*, the data used to fit the model, though not necessarily on new data. This is known as *overtfitting* because the model tends to fit the training data precisely but cannot generalize when presented with new data.

The problem of choosing the best subset of predictors, e.g., those that minimize prediction error, is known as *variable selection*. It is worth noting that, since five of the eight predictors in Table 1 are *categorical*, the number of effective predictors used to estimate *f* is much larger than eight.[4] Rather than considering all possible combinations of the predictors (sometimes called stepwise variable selection), this paper uses *regularization* to find the minimal model, $\hat{f}_{minimal}$. Regularization penalizes model complexity (i.e., increasing the number of predictors) while minimizing a criterion such as prediction error (Zou and Hastie, 2005). Regularization is often used to prevent overfitting and can be used to enforce model simplicity.

In particular, the paper uses the lasso (least absolute shrinkage and selection operator) to estimate $\hat{f}_{minimal}$.[5] Lasso is especially useful for obtaining *sparse* solutions: the larger the penalty, the lower the number of *active* predictors used to train the model. Thus, regularization via lasso performs variable selection.

Not all regularization methods perform variable selection. For instance, ridge regression (Hoerl and Kennard, 1970) reduces the influence of correlated groups of predictors, but ultimately all predictors are used to train the model.[6] Lasso, on the other hand, does not account for correlations among predictors. The *elastic net* is a

[3] K-fold cross-validation splits the data into K mutually exclusive subsets then trains the model K times. At each iteration, one subset is used to estimate prediction error while the rest of the data is used to estimate $\hat{f}$. See Krstajic et al. (2014) for a review of the procedure as well as discussion of the potential pitfalls.

[4] In fact, Fung et al. (2018b) also include the combined effect of seismicity and the performance objective, an *interaction term*, that brings the total number of effective predictors to 35.

[5] Lasso penalizes the sum of the absolute values of the coefficients, $\sum_k|\beta_k|$, i.e., the L_1-norm of the coefficient vector, $\|\beta\|_1$ (Tibshirani, 1996).

[6] Ridge regression penalizes the sum of the squared values of the coefficients, $\sum_k {\beta_k}^2$, i.e., the squared L_2-norm of the

regularization method that combines lasso with ridge regression, thus performing variable selection while accounting for correlations among predictors (Zou and Hastie, 2005). Elastic-net regularization is a convex combination of lasso and ridge regression and therefore includes both as special cases.

In order for $\hat{f}_{minimal}$ (and $\hat{f}_{practical}$) to be compared against an appropriate benchmark, the benchmark model $\hat{f}_{benchmark}$ should be chosen as the optimal form of elastic-net regularization, which includes as a special case a model with no penalty (and hence no regularization). The model selection step of nested K-fold cross-validation is used to choose $\hat{f}_{benchmark}$.[7] The model evaluation step of nested K-fold cross-validation is then used to estimate the expected out-of-sample performance of each model.

The Training Data

The historical retrofit cost data used in this paper was originally collected for FEMA 156 (FEMA, 1994) and is freely available online. In particular, the data can be found as part of FEMA's archived Seismic Rehabilitation Cost Estimator (SRCE) software, (FEMA, 2013–2014).

The publicly available version of the data (the SRCE data) includes 1978 buildings, compared to the 2088 collected for FEMA 156. The SRCE data set is missing an important building characteristic that is used in FEMA 156: building occupancy class. Nevertheless, the discussion in FEMA 156 suggests this data set should be representative of commercial and residential buildings in the United States and Canada.

The SRCE data used for training includes all of the predictors shown in Table 1, as well as the structural seismic retrofit cost for each building. Building model types are categorized into eight *building groups*, as shown in Table 2.

Table 2. Building types, building groups, and their shares in the SRCE data.

Building Group	Building Type	Building Type Name	Share
1	URM	Unreinforced Masonry	30.08 %
2	W1	Wood Light Frame	2.62 %
	W2	Wood (commercial or industrial)	3.08 %
3	PC1	Precast Concrete Tilt Up Walls	3.34 %
	RM1	Reinforced Masonry with Metal or Wood Diaphragm	3.34 %
4	C1	Concrete Moment Frame	6.75 %
	C3	Concrete Frame with Infill Walls	16.64 %
5	S1	Steel Moment Frame	4.85 %
6	S2	Steel Braced Frame	1.83 %
	S3	Steel Light Frame	0.72 %
7	S5	Steel Frame with Infill Walls	7.01 %
8	C2	Concrete Shear Wall	16.19 %
	PC2	Precast Concrete Frame with Infill Walls	0.79 %
	RM2	Reinforced Masonry with Precast Concrete Diaphragm	0.66 %
	S4	Steel Frame with Concrete Walls	2.10 %

Table 3 presents a summary of statistical information from the SRCE data for the structural retrofit cost per square foot, as well as some of the predictors from Table 1. All cost and RMSE values in this paper are given in 2016 US dollars per square foot (1 ft = 0.3048 m). Further information on the data set can be found in Fung et al. (2018b).

coefficient vector, $\|\beta\|_2^2$ (Zou and Hastie, 2005).

[7] To be precise, α, the hyperparameter governing the convex combination of L_1 and L_2 penalties, is chosen using random grid search. For a given α, the penalty parameter, λ, is chosen based on search algorithms developed in Friedman et al. (2010). If $\lambda = 0$, then there is no penalty and thus no regularization. The best combination of α, λ is the best pair (in terms of RMSE) from the model selection step of nested K-fold cross-validation.

Table 3. Summary statistics for the outcome of interest, structural retrofit cost per square foot (1 ft = 0.3048 m) and select predictors in the training (SRCE) data, with N=1526 excluding Canadian buildings.

Variable	Minimum	Mean	Median	Maximum	Standard Deviation
Structural cost ($/sq ft)	0.49	36.03	23.33	675.42	44.74
Area (1000 sq ft)	0.15	68.98	28.67	1430.30	113.26
Age	2.00	44.29	40.00	153.00	22.13
Stories	1.00	3.12	2.00	38.00	2.99

RESULTS

This section presents the main results. First, the results from model selection and the resulting benchmark model. Second, the expected out-of-sample performance for the benchmark, minimal, and practical models.

Model Selection and the Benchmark Model

The benchmark model is the optimal model chosen in the model selection step: it is the model that minimizes prediction error in the model selection step of nested K-fold cross-validation. Table 4 presents the results of model selection and, hence, the resulting benchmark model. The value of $\alpha = 0.7$ implies elastic-net regularization that weighs lasso more heavily than ridge regression and, thus, favors a sparser model. The value of $\lambda = 0.002$, while not huge in magnitude, is sufficiently larger than zero that the regularization penalty is non-trivial.

Table 4. The benchmark model is defined as the model with optimal values of the regularization hyperparameters, α, λ based on nested K-fold cross-validation with K=10.

	$\boldsymbol{\alpha}$	$\boldsymbol{\lambda}$
Mean	0.707	0.002
Standard deviation	0.230	0.001

Table 5 presents prediction error results from the model selection step of nested K-fold cross-validation for each of the candidate models. It is worth noting that Fung et al. (2018b) do not use regularization. Since the optimal model from the model selection step, $\hat{f}_{benchmark}$, does use some form of regularization, it is worth comparing the benchmark model to the original model with no regularization, $\hat{f}_{original}$.

Table 5. Model selection for each of the candidate models based on prediction error, RMSE, as well as the standard deviation of RMSE, σ_{RMSE}. The benchmark model is defined as the optimal model: the model with the lowest RMSE. All values in dollars per square foot (1 ft = 0.3048 m).

Model	RMSE	σ_{RMSE}
Benchmark model, $\hat{f}_{benchmark}$	38.57	1.06
Original model, $\hat{f}_{original}$	38.89	1.06
Minimal model, $\hat{f}_{minimal}$	38.80	1.17
Practical model, $\hat{f}_{practical}$	40.11	1.06

The results suggest that the practical model, $\hat{f}_{practical}$, is not optimal. This is not surprising: if a decision maker has all of the information necessary for prediction available, it should be used. Regularization will ensure that the model does not overfit. Nevertheless, the standard deviation of the RMSE estimates suggest that the practical model is only marginally sub-optimal.

The results suggest that some form of regularization is optimal and, in particular, regularization that favors sparsity. Note that the minimal and original models are only marginally sub-optimal (in terms of RMSE) to the benchmark model. In fact, the standard deviation of the RMSE estimates suggests that both $\hat{f}_{minimal}$ and $\hat{f}_{original}$ could be optimal. Thus, a decision maker could reasonably choose either $\hat{f}_{benchmark}$, $\hat{f}_{minimal}$, or $\hat{f}_{original}$.

On the other hand, the minimal model $\hat{f}_{minimal}$ uses 24 of 35 predictors and is thus only marginally sparser than the benchmark model $\hat{f}_{benchmark}$, which uses 25 of 35 predictors. In both cases, the models ignore several of the interactions between seismicity and the performance objective, suggesting that the other predictors do a good job of capturing this effect. Interestingly, both models discard the effect of the Damage Control performance objective, most likely because its effect on cost is correlated with the Life Safety and Immediate Occupancy performance objectives. The most noteworthy result of shrinkage is that the effect of building area as a predictor of cost is on the order of 10^{-6}, almost (though not identically) zero.

Model Performance

This section presents the estimates of expected out-of-sample performance for each of the candidate models (including the original model with no regularization). Expected out-of-sample performance is an estimate of prediction error on *new data*, obtained from the model evaluation step. The results are presented in Table 6.

Table 6. Estimated out-of-sample performance, RMSE, and standard deviation of RMSE, σ_{RMSE}, for each of the candidate models. All values in dollars per square foot (1 ft = 0.3048 m).

Model	RMSE	σ_{RMSE}
Benchmark model, $\hat{f}_{benchmark}$	38.61	8.97
Original model, $\hat{f}_{original}$	37.72	8.91
Minimal model, $\hat{f}_{minimal}$	37.69	8.91
Practical model, $\hat{f}_{practical}$	39.15	8.15

The main implication is that the minimal model, $\hat{f}_{minimal}$, dominates the benchmark model, $\hat{f}_{benchmark}$, in terms of performance on new data. Indeed, the minimal model outperforms all the candidate models, including the original model, $\hat{f}_{original}$. However, note that the standard deviation of RMSE estimates is much larger when estimating out-of-sample performance than when selecting a model (as shown in Table 5). Thus, the differences between RMSE estimates presented in Table 6 are not statistically significant, meaning the minimal model is no worse than the benchmark model.[8]

Most importantly, note that while the practical model, $\hat{f}_{practical}$, has the worst expected performance (as shown by the largest RMSE value in Table 6), the large variance in prediction error estimates suggests the practical model could achieve performance comparable to the benchmark or minimal models. In other words, the results suggest that there is no statistically significant penalty to deliberately omitting (or not collecting) information on building age, building height, and building type.

[8] For instance, application of Welch's *t*-test shows the difference in RMSE estimates for the benchmark and minimal models has a *p*-value of 0.823 and, thus, is far from statistically significant.

CONCLUSIONS

This paper considers an "optimal" model for predicting structural seismic retrofit costs and considers the impact on expected out-of-sample performance (prediction error on new data) when the model does not include all of the available predictors. The results suggest that, in fact, the optimal model itself does not require all of the predictors. Moreover, the paper considers a model that deliberately omits information on building age, building height, and building model type, as a thought experiment on what happens when a decision maker cannot (or does not want to) obtain this information for a portfolio of buildings. The results suggest there is a negligible penalty for omitting this information from the model. In fact, the results suggest all models achieve statistically indistinguishable performance on new data.

It would be interesting to compare performance of the GLM predictive model used in this paper with other models that can capture more nonlinearities (e.g., random forests, gradient boosting machines, and deep neural networks). This is left for future work.

Disclaimer

NIST policy is to use the International System of Units (metric units) in all its publications. In this paper, however, information is presented in U.S. Customary Units (inch-pound), as this is the preferred system of units in the U.S. earthquake engineering industry.

REFERENCES

FEMA. Typical Costs for Seismic Rehabilitation of Existing Buildings, Volume 1: Summary. *FEMA 156*, Federal Emergency Management Agency, 1994.

FEMA. SRCE: Seismic Rehabilitation Cost Estimator. https://www.fema.gov/media-library/assets/documents/30220, 2013–2014.

Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 2010, 33:1 1-22.

Fung JF, Butry DT, Sattar S, McCabe S. A Methodology for Estimating Seismic Retrofit Costs. *National Institute of Standards and Technology Technical Note*, 2017, NIST TN 1973.

Fung JF, Sattar S, Butry DT, McCabe S. Cost Estimates for the Seismic Retrofit of Federally Owned and Leased Buildings. *11th National Conference on Earthquake Engineering*, Los Angeles, 2018a.

Fung JF, Sattar S, Butry DT, McCabe S. A Predictive Modeling Approach to Estimating Seismic Retrofit Cots. *Earthquake Spectra* (under review), 2018b.

Fung JF, Sattar S, Butry DT, McCabe S. Estimating Structural Seismic Retrofit Costs for Federal Buildings. *National Institute of Standards and Technology Technical Note*, 2018c, NIST TN 1996.

Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 1970, 12:1 55-67.

Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 2014, 6:10.

Moran JL, Solomon PJ, Peisach AR, Martin J. New Models for Old Questions: Generalized Linear Models for Cost Prediction. *Journal of Evaluation in Clinical Practice*, 13: 381-389.

Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 1996, 58: 267-288.

Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 2005, 67: 301-320.

Proceedings of the ASME 2019 14th International Manufacturing Science and Engineering Conference
MSEC2019
June 10-14, 2019, Erie, PA, USA

MSEC2019-2783

AN ANALYSIS OF RECENT STANDARDS ON COMPOSITE PRODUCT MODELS TO ENABLE DIGITAL TRANSFORMATION OF COMPOSITE PRODUCT MANUFACTURING

Thomas Hedberg, Jr.[1], Allison Barnard Feeney, Vijay Srinivasan
Engineering Laboratory, National Institute of Standards and Technology
Gaithersburg, MD 20899, U.S.A.

Keith Hunten, P.E.
Lockheed Martin Corporation (Retired)
Fort Worth, TX 76108, U.S.A.

ABSTRACT

ASME and ISO are issuing new editions of their standards that deal with definitions and models for composite products. The new edition of ASME Y14.37 deals with standardized product definition for composite parts. The Edition 2 of ISO 10303-242 contains standardized data models for three-dimensional representations of composite products. This paper analyzes several salient features of these two standards with a focus on their potential impact on the digital transformation of composite product manufacturing. It also provides a mathematical and information theoretic exposition of *ply table* as an important representation for modeling some of the complex composite products and their manufacturing processes.

1. INTRODUCTION

Products made from composite materials can be found everywhere, from recreational sports (e.g., tennis racket) to advanced transportation (e.g., aircraft structure). Composite materials, which are hybrids of different materials, fill some crucial holes in the material-property space that are left open by conventional monolithic materials. When stiff, strong, tough, and light materials are needed, the designers often turn to a hybrid of materials, and to novel manufacturing processes to combine these materials to achieve their design objectives [1].

There are many ways to combine two or more materials to create a composite product. Among such products are *fibrous composites*, which consist of reinforcement fibers embedded in a plastic matrix such that the fibers do not separate from the matrix when the composite product is loaded. The fibers may be made up of materials such as glass, carbon, or aramid (e.g., Kevlar). The matrix may consist of thermosetting resins such as polyester or epoxy; sometimes thermoplastic resins may be used as the matrix. The fibers may be continuous strands or chopped up into smaller pieces before they are embedded in the matrix resin. The composite product may also contain *sandwich cores* that are lighter. This paper deals with composite products that are made from fiber-reinforced plastics (FRP) and may have sandwich cores.

Recent interest in energy efficient products and renewable energy production has accelerated the use of composite products [2]. FRP is used to reduce the weight of planes and cars, thereby contributing to fuel efficiency. Large turbine blades for windmills are made almost exclusively from FRP. Renewable energy sources (such as wind, waves, and solar) are notoriously intermittent and require energy storage. Flywheels and compressed air tanks are made using FRP, and these composite products provide the mechanical and pneumatic means, respectively, for storing such intermittently harvested energy. All these applications are made possible by the light weight, high strength, and resistant to corrosion offered by FRP.

As FRP products are gaining popularity, their manufacturing processes are attracting greater attention. The FRP design and manufacturing have remained in the hands of highly skilled engineers for several decades. During this period, several innovative design and manufacturing practices developed by these engineers have led to the introduction of numerous successful products to the market. However, many of these practices have remained *ad hoc*, and a lack of their systematization and standardization has hindered the integration of engineering information systems used in the composite product design and manufacturing processes. This problem has now become even more acute as the manufacturing industry is going through a digital transformation, a phenomenon that is variously called smart manufacturing [3], cyber-manufacturing [4], and Industrie 4.0 [5, 6].

[1] Corresponding author: tdh1@nist.gov

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay.
"AN ANALYSIS OF RECENT STANDARDS ON COMPOSITE PRODUCT MODELS TO ENABLE DIGITAL TRANSFORMATION OF COMPOSITE PRODUCT MANUFACTURING."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

Standards development organizations such as ASME and ISO (International Organization for Standardization) have responded to urgent calls to enable this digital transformation and have been actively producing standards to serve the composites manufacturing industry. Two such standards have been revised for industrial use recently. One is a new edition of the ASME Y14.37 standard that deals with standardized product definition for composite parts [7]. The other is Edition 2 of the ISO 10303-242 standard that contains standardized data models for three-dimensional representations of composite products [8].

This paper analyzes several salient features of these two standards with a focus on how they may enable the digital transformation of composite product manufacturing. A major technical contribution of this paper is the mathematical and information theoretic exposition of what is called a *ply table* as an important representation for modeling some of the composite products and their manufacturing processes.

The rest of the paper is organized as follows. Section 2 provides a brief introduction to information models needed for design and manufacturing of composite products. Section 3 introduces the new edition of the ASME Y14.35-2019 standard on product definition for composite parts. Section 4 provides a brief analysis of the ISO 10303-242:2019 (also known as STEP AP 242) standard that includes composite products. Some of the future directions for research and standardization are discussed in Section 5. The paper concludes with a summary in Section 6.

2. COMPOSITE PRODUCT INFORMATION MODELS

This paper focuses on thin-walled FRP products. The design and manufacturing are very closely tied to each other for such products. In this sense, FRP products are very similar to semiconductor chips; both are led by innovations in manufacturing processes and both are built as layers involving multiple materials, with one important difference. In the case of FRP, plies and laminates can be laid up on curved surfaces and sandwich cores to produce complex three-dimensional structures. This crucial difference from semiconductor chips, which are built predominantly of planar layers, introduces some important geometrical challenges in the information modeling of composite product structures.

Figure 1 illustrates a general, high-level classification of manufacturing processes employed to produce FRP products [9, 10]. It is not intended to be the final word in the classification of FRP manufacturing processes because new technologies are being introduced at a rapid pace, thus constantly changing the manufacturing landscape. However, the classification of Fig. 1 is sufficiently general and inclusive to guide the information modeling technology and related standardization processes.

As indicated in Fig. 1, industry employs both thermosetting and thermoplastic composites. Thermosetting plastics are more popular, but they are not recyclable because of irreversible polymerization of the resins used in the manufacturing process. There is an increasing interest in replacing thermosetting plastics with thermoplastics that are recyclable and hence eco-friendlier.

Both short and continuous fibers are used in composite products, as indicated in Fig. 1. While short fibers are cheaper and easier to handle, industry prefers continuous fibers when high performance in strength and stiffness are needed (for example, in aerospace applications). This is the primary reason for the importance given to continuous fibers in the ASME and ISO standards that are addressed in Sections 3 and 4.

All composites manufacturing processes described in Fig. 1 depend on a curing process, which involves application of temperature and pressure, on a finite stack (of layers of resins and fibers) to obtain the final product. Some resins do not require additional temperature or pressure than the ambient conditions for curing. Curing is a chemical process that enables polymers in the matrix resins to cross-link, which produces a harder and more homogeneous matrix within which the fibers (both short and continuous) are firmly embedded. The curing process will introduce changes, both noticeable and invisible, in the product. Therefore, it is important to distinguish between the 'cured' and 'uncured' states of a composite structure. As described in Sections 3 and 4, standards pay greater attention to the uncured state of the composite structure for reasons that will be explained during this paper.

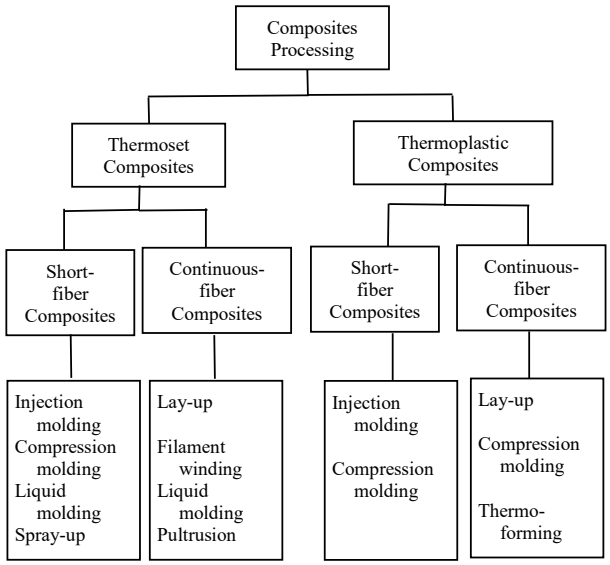


FIGURE 1: Classification of FRP manufacturing processes [9].

With these preliminaries, it is now possible to identify the following set of elements and entities that are common to all FRP products, and therefore candidates for standardization.

- *Tools and molds*: All composite processes identified in Fig. 1 start with tool surfaces (e.g., for spray-up and lay-up) and molds. Similar to casting and injection molding, the shape and accuracy of an FRP product depends critically on these tool and mold surfaces and the solids bounded by these surfaces. The standards must provide means to represent these three-dimensional surfaces and solids. Luckily,

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay.
"AN ANALYSIS OF RECENT STANDARDS ON COMPOSITE PRODUCT MODELS TO ENABLE DIGITAL TRANSFORMATION OF COMPOSITE PRODUCT MANUFACTURING."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

standards for such geometric representations are already available because they are needed for conventional products that are routinely manufactured every day [11, 12].

- *Fibers*: Fibers are quite correctly called the 'reinforcements' in composite products. Fibers are the load-bearing members and are the major contributors to the strength and stiffness. Fibers have a circular cross-section and can be made up of materials such as carbon, glass, or aramid (e.g., Kevlar). A typical composite product may contain thousands – often millions – of fibers even when the fibers are continuous; short fibers make the count even larger. This poses a serious problem in geometrical modeling of these important elements in FRP products.

  Luckily, continuous fibers are placed in a mathematical pattern within each ply and these plies are arranged (in an uncured state) geometrically to form laminates, which may be arranged further in a well-defined manner to form the composite product. Even when short fibers are used, they are usually placed randomly within a ply and this randomness provides a modeling abstraction for standardization. In any case, the hierarchical description of plies, laminates, and the uncured state of the final product provides a mathematical means to conquer the size complexity of millions of fibers. This then enables an information theoretic abstraction for standardization, as described in Section 3 and 4.

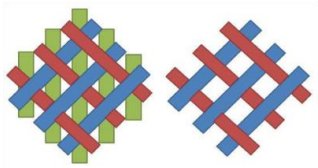
- *Resins*: Resins provide the medium, often called the matrix, within which the fibers are embedded. The resins transfer loads among the fibers and they provide much needed protection to the fibers from ambient environment (e.g., resistance to corrosion). They are also responsible for the ductility and toughness of the composite product.

  Thermosetting resins include epoxy and polyester. When subjected to temperature and pressure (sometimes even under the ambient conditions), polymers in these thermosetting resins form cross-links and result in a harder substance. This chemical process is irreversible, and this renders these resins non-recyclable. Thermoplastic resins do not suffer from this drawback, but such resins are being developed only recently for large-scale industrial use.

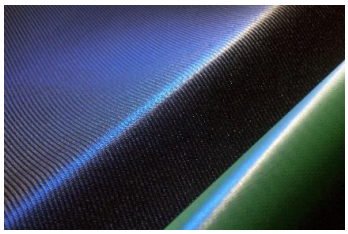
  From an information modeling perspective, resins can be viewed as homogeneous and isotropic materials. But resins are never used in isolation in FRP products. Reinforcement fibers are embedded in these resins and these fibers provide the desired non-homogeneity and anisotropy. This leads to geometric modeling challenges that will be addressed below.

- *Plies*: A ply is usually an arrangement of reinforcement fibers in a resin matrix. More formally, a ply is 'one discrete piece of manufactured material (e.g., fabric, tape, adhesive film)' [7], thus generalizing the ply definition to include adhesive films as well. This formal definition elucidates the critical fact that a ply is an important *discrete* module in an FRP product, and therefore it is a basic entity in modularizing the design and manufacturing of composite products. Thus, plies play a dominant role in the uncured state of an FRP product.

  A ply can include short or continuous fibers. Dry continuous fibers can be spun, braided, woven, or stitched using any number of textile processes to produce a 'preform.' Figure 2(a) shows some preform patterns, with thousands of fibers in each strip. These preforms can then be injected with wet resins during the manufacturing process. Alternatively, such textile or plain fibers can be impregnated with resins to produce a 'prepreg' that can be subjected to curing in a later manufacturing process. Figure 2(b) shows a prepreg fabric that can be used as a ply, after it is trimmed as needed.



(a)



(b)

FIGURE 2: Examples of (a) Preform and (b) Prepreg in creating plies [13, 14].

The plies will be stacked up in layers, starting from a tooling surface. As mentioned earlier, each ply can contain thousands – sometimes millions – of fibers. A pragmatic approach, which is probably the only sensible approach, adopted by industry is to model each ply in it uncured state as a combination of four sets of information: (1) The volume occupied by the ply as a homogeneous and isotropic three-dimensional solid, (2) The sequence in which the ply is laid up during manufacturing, (3) Arrangement of fibers within the ply and relative to the ply (or tool surface) before it in the sequence, with particular attention to fiber orientation, and (4) Information about the fiber material, resin material, and geometrical pattern of fibers used in the ply preparation.

- *Laminates*: A laminate results from the stack up of two or more plies in an uncured state, as illustrated using a simple example in Fig. 3. When cured, the resins in the plies link up by polymerization to form the composite product as

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay.
"AN ANALYSIS OF RECENT STANDARDS ON COMPOSITE PRODUCT MODELS TO ENABLE DIGITAL TRANSFORMATION OF COMPOSITE PRODUCT MANUFACTURING."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

shown in Fig. 4. The interfaces between plies have disappeared in Fig. 4, leaving only the fibers that are now embedded in one cured conglomerate of resin matrix.

The ply sequence in a laminate and the fiber orientations in the plies in a laminate are two of the most important pieces of information in engineering practice. This can be mathematically formalized by postulating the following invariance properties under nominal (or ideal) curing operations.

1. Preservation of three-dimensional arrangement (i.e., the order and sequence in three-dimensions) of fibers. This can be viewed as the invariance of the combinatorial topology of the fibers in a laminate before curing (e.g., in uncured state as in Fig. 3) and after curing operation (e.g., in cured state as in Fig. 4).

2. Preservation of orientation of fibers. This can be viewed as the invariance of angles between fibers before curing (e.g., in uncured state as in Fig. 3) and after curing operation (e.g., in cured state as in Fig. 4) that may result in uniform scaling due to volumetric shrinkage.

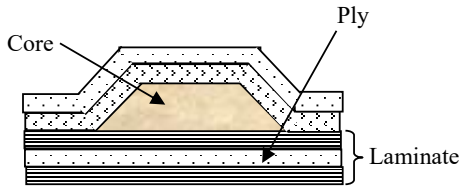


FIGURE 3: Illustration of uncured plies, laminates, and core.

Under actual (as opposed to nominal or ideal) lay-up and curing operations, the invariance of combinatorial topology may still be preserved; but, some small changes in the relative orientation of fibers should be expected. The distances between fibers in adjacent plies may, however, undergo larger changes during curing.

- *Sandwich cores*: A core is a lightweight component sandwiched between laminates and bonded to the laminates. A primary role of such a core is to increase the section modulus of thin-walled structures without increasing their weight considerably [1]. Figure 3 illustrates a cross-sectional view that includes plies, laminates, and a core.

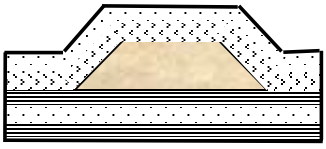


FIGURE 4: Illustration of composite product in cured state.

The design philosophy that underpins the development of composite products follows the classical causal links of processing, structure, property, and performance as shown in Fig. 5 [15]. In analyzing the performance of a composite product, engineers start with the manufacturing processes adopted for that product. Some of these processes are captured in the information modeling of plies, laminates, and cores as described thus far. But these provide a model for only the uncured state (e.g., as in Fig. 3) of the composite structure. From this, a model for the cured state of the composite (e.g., as in Fig. 4) is obtained by applying a combination of computational techniques and empirical rules derived from experiments (for example, to account for shrinkage).

The structure thus obtained is then subjected to computational analysis (e.g., finite element analysis) to predict its properties and performance under various service conditions. A typical composite product design is an iterative process, involving a cycle implied in Fig. 5 that shows a causal progression in one direction and a synthesis progression (starting with the performance goal) in the other direction.

The brief exposition of composite product information model presented in this section forms the basis for analyzing two recent standards. Section 3 focuses on the recent revision of the ASME Y14.37 standard, followed by Section 4 that addresses Edition 2 of the ISO 10303-242 standard.

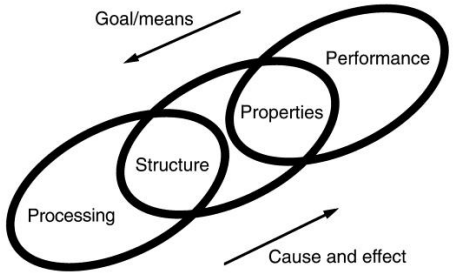


FIGURE 5: The causal links of processing, structure, properties, and performance [15].

3. ASME Y14.37-2019 STANDARD

ASME issued its first standard on composite parts in 2012, when it was called 'Composite Part Drawings' [16]. It reflected the practice at that time to focus on two-dimensional drawings as the primary means of defining engineering products. While revising this standard, the title was changed to 'Product Definition for Composite Parts' to acknowledge the increasing industrial use of three-dimensional models and machine-readable representations of products and manufacturing processes [7]. This new edition also started harmonizing its definitions with other standards, such as the ISO STEP standards [8], which deal with three-dimensional data models to represent products that include FRP products.

There are other ASME standards that deal with dimensioning and tolerancing drawings and three-dimensional models [11, 17]. These can be used for composite products as well, but these standards are not sufficient due to the complexity

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay.
"AN ANALYSIS OF RECENT STANDARDS ON COMPOSITE PRODUCT MODELS TO ENABLE DIGITAL TRANSFORMATION OF COMPOSITE PRODUCT MANUFACTURING."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

of FRP structures. To address this deficiency, ASME Y14.37-2019 deals with those composites definitions that are not covered by the other ASME Y14 standards.

Broadly speaking, the types of information standardized by ASME Y14.37-2019 fall under two categories. The first category is the geometric information, either as two-dimensional drawings (projected views and cross-sections) or three-dimensional models, which are already covered by the above-mentioned ASME standards [11, 17]. The second category relies heavily upon *ply tables*, which are unique to composite products. The ply tables also refer to the geometric information mentioned in the first category to define the geometry of plies, cores etc. Thus, the ply tables and the geometric information complement each other for composite products.

Since ply tables play such an important role for FRP products, the rest of this section is devoted to them. A typical ply table, in its simplest form, is shown in Table 1. Such a ply table is accompanied by a three-dimensional model, or a two-dimensional presentation such as the one shown in Fig. 6. The ply table alone will not capture the composite product model; it is the combination of the ply table and the accompanying geometric model (or drawing) that conveys a more complete information.

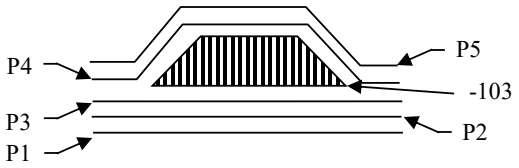
TABLE 1: A simple ply table.

-101 ASSEMBLY			
PLY LEVEL	PLY/ITEM	ORIENTATION	MATERIAL
1	P1	0°	10745
2	P2	90°	10721
3	P3	0°	10745
4	-103 CORE		
5	P4	45°	10679
6	P5	90°	10721

The rows and columns of Table 1, and the accompanying Fig. 6, require some explanation. Each ply in Fig. 6 is graphically presented as a curve (with straight line segments, in this case) without any thickness attribute. In three-dimensions, a ply will be presented as a surface without thickness. The first row in Table 1 identifies an assembly of plies as -101, which is shown in Fig. 6, and this assembly is also called a laminate. The first column specifies the ordered sequence in which the plies and the core are arranged in the laminate (which is identified as the assembly -101). The plies and core are then identified as named items in the second column and are thus labeled in the geometrical presentation of Fig. 6. The core is identified and labeled as -103 in Table 1 and in Fig. 6, respectively.

As described in Section 2, the ply level in the first column of Table 1 provides a representation of the combinatorial topology of the fibers in the laminate, and this topology remains invariant under the curing operation. The ply level also provides the sequence in which the plies (and the core) should be stacked to form the laminate. Thus, it also serves as an important part of the FRP manufacturing process specification.

The third column in Table 1 specifies the orientation of the fibers in each ply with respect to a reference system, which may be fixed on a tool surface. This is a representation of another



-101 ASSEMBLY

FIGURE 6: Geometrical companion to the ply table in Table 1.

invariant geometric relationship (i.e., angles between fibers) that is preserved under the nominal (or ideal) curing operation with uniform volumetric shrinkage, as discussed in Section 2. The fourth column identifies the material associated with each ply. Such material identification can refer to a much richer definition of each ply, which is not covered in the ASME standard. The ply thickness (in uncured state) may also be defined under the material category; the thickness of the FRP product after curing – also known as 'consolidated thickness' – may differ from the sum of the ply thicknesses due to volumetric shrinkage.

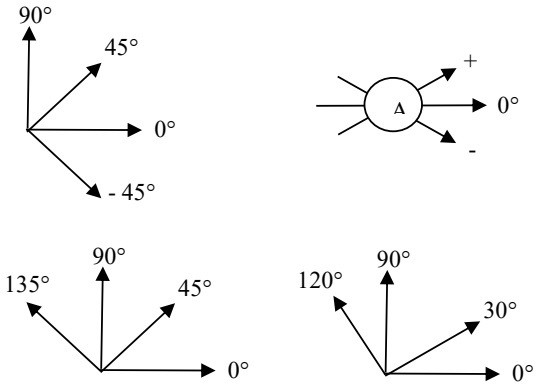


FIGURE 7: Examples of ply orientation symbols.

To support the orientation specification in column 3 of Table 1 for complex three-dimensional products, a more elaborate definition may be necessary. This is accomplished using an object called a *rosette* and an orientation symbol associated with that rosette. Figure 7 shows some examples of orientation symbols, which look like the needles on a two-dimensional magnetic compass used to get a local bearing on direction while standing on an undulating terrain – these symbols serve a similar purpose to orient fibers in a ply on a curved surface.

Once a ply orientation symbol is chosen from Fig. 7 and is associated with a rosette, it defines a two-dimensional coordinate system – either a Cartesian or a polar coordinate system. Using the right-handed rule, a unique normal to the two-dimensional plane of the orientation symbol can be found, thus forming a full-

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay.
"AN ANALYSIS OF RECENT STANDARDS ON COMPOSITE PRODUCT MODELS TO ENABLE DIGITAL TRANSFORMATION OF COMPOSITE PRODUCT MANUFACTURING."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

fledged three-dimensional coordinate system. The origin of such a coordinate system associated with a rosette can be initially place at a specific point on a ply surface, and its coordinate axes can be appropriately oriented to indicate the fiber orientation in the ply at that initial point.

In some flat FRP products a single rosette per ply might suffice. But in more complex products, especially those that involve non-developable surfaces, more than one rosette per ply will be necessary. To facilitate the specification of these additional rosettes in a ply, ASME Y14.37-2019 defines five types of transformation rules, which transform the initial rosette on a ply to any or specific points on the ply surface. Figure 8 illustrates one such type (Type 2, in this case) of transformation that can be specified to guide the lay-up (also, to guide wrapping and draping of 'cloths') of the ply so that the 0° fiber orientation is along the indicated guide curve. Other transform types, including a user defined type, are among the five ply orientation transformation types defined in ASME Y14.37-2019.

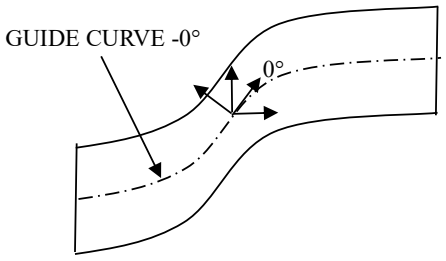


FIGURE 8: Type 2 rosette transform per ASME Y14.37-2019.

Even beyond the rosettes, some FRP products will require more complex information than those provided by the likes of Table 1 and Fig. 6. ASME Y14.37-2019 provides means to add rows and columns to the ply table beyond the simple example shown in Table 1. For example, when pultrusion is used to produce FRP products, the ply table is expanded to cover other information types such as roving, ply count, ply yield, and percentage weight.

The brief analysis of the ASME Y14.37-2019 standard presented in this section captures only the important features to illustrate the progress that has been made thus far. It has focused on ply tables and their geometric companions. More detailed information can be found in the ASME standard document [7].

4. ISO 10303-242:2019 STANDARD

ISO 10303, commonly known as STEP (Standard for Exchange of Product model data), is a family of international standards designed to exchange digital information of engineered products, enabling an ever-widening range of engineering information systems to interoperate [18–25]. An important recent member of the STEP family is *ISO 10303-242, Application protocol: Managed model-based 3D engineering* (commonly referred to as AP 242), which has quickly emerged as a critical enabler of digitization of manufacturing [26].

The first STEP standard for representing composite shape and structure was published in 2001 as *ISO 10303-209:2001, Application protocol: Composite and metallic structural analysis and related design* (commonly referred to as AP 209). In this standard, composite structure definitions were integrated with both configuration-controlled three-dimensional design and finite element analysis disciplines [27]. In preparation for the second edition of AP 242 and the third edition of AP 209, significant changes have been made to STEP data models for three-dimensional representations of composite structures, particularly for ply orientation (rosettes) and ply tables. Recent work on rosettes has been coordinated between the ISO subcommittee on Industrial Data and the ASME subcommittee Y14.37, whose standard was described in Section 3. In keeping with these new developments, significant changes to the analysis domain are also being planned (only in AP 209 Edition 3).

The APs 242 and 209 are specified according to a STEP modular architecture [28]. Both APs include the same three-dimensional composite structure representation. In addition, AP 209 supports the causal links of processing, structure, properties, and performance depicted in Fig. 4. So, AP 209 supports a design product structure and an analysis product structure, with versioning of each and relationships between the design and analysis models. This means that there may be representations of a nominal design shape, analysis shape(s), and optimized analysis shape(s). Any healing or meshing based on the nominal geometry may be stored separately with links to the nominal geometry being modified.

The STEP data model for composites is divided into the following five parts: (1) Part and zone laminate tables [29], (2) Composite constituent shape [30], (3) Composite constituent material aspects [31], (4) Stock material [32], and (5) Ply orientation specification [33]. Each of these parts will be explained briefly below.

Part and zone laminate tables, whose data model is shown in Fig. 9, is the core specification of the STEP composite structure. Note that what the ASME Y14.37-2019 calls a 'ply table' is referred to as a 'laminate table' in Fig. 9. A closer examination reveals entity definitions for base surface, direction for material lay-up, ply orientation, and ply sequence. Also included are information for the material properties, ply thickness, and volume percentage of fibers in the ply; these are important for the finite element analysis of the composite product.

A more detailed model for the shape and material of the ply, core, resin, and fiber is shown in Fig. 10. It defines composite constituents such as processed core, ply orientation, ply sequence, and woven and braided fiber filaments. It also refers to the stock materials from which the core, resins, and fibers are made. A separate stock material module defines the material stock of composite constituents and how they are approved and shaped; it also defines their material aspects and versions [32]. The composite constituent placement, net shapes, and boundaries are described by shape representations such as *Advanced_brep, Edge_based_wireframe*, *Faceted_brep*, *Shell_based_wireframe*, *Tessellated_shape*, and *Three_d_geometry_set*.

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay.
"AN ANALYSIS OF RECENT STANDARDS ON COMPOSITE PRODUCT MODELS TO ENABLE DIGITAL TRANSFORMATION OF COMPOSITE PRODUCT MANUFACTURING."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

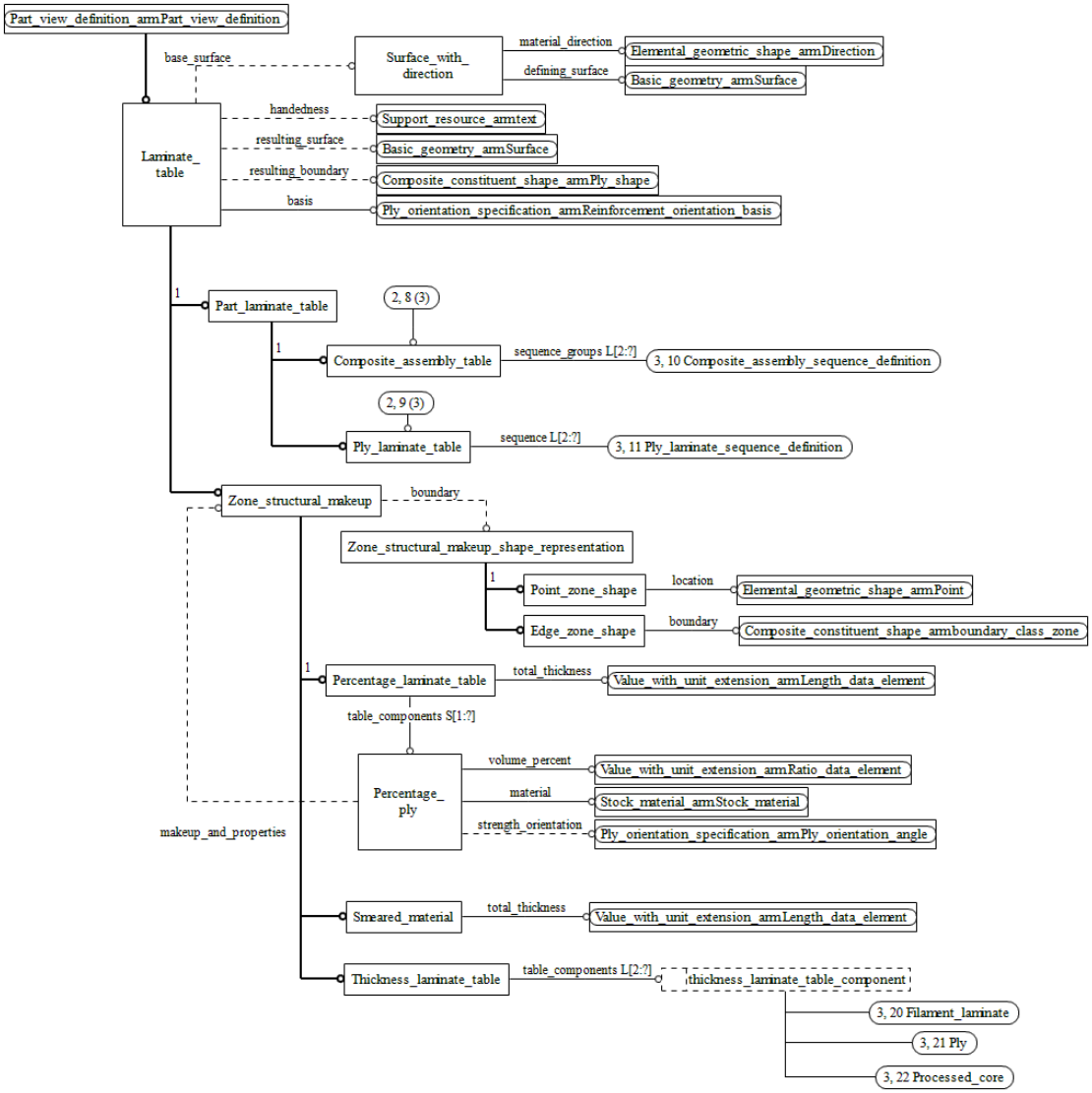


FIGURE 9: STEP data model for laminate table [29].

Figure 11 shows a more detailed data model for ply orientation, with emphasis on the orientation of reinforcement fibers. As described in Section 3, several types of rosettes are used for this purpose. Figure 11 defines many such fiber orientation types, including guide curve that is shown in a simple example in Fig. 8. Definition of these rosettes have been harmonized with the ASME Y14.37-2019 standard.

The new editions of AP 242 and AP 209 also make it possible to identify material specifications from internal and external document references, and properties for specific operating environments. The current best practice for composite product definition is to define the uncured (i.e., pre-autoclave) components as a typical assembly and then use a derived ('made from') solid to represent the cured part.

Such a cured solid model could then be defined as a 'make from' part that consists of the inseparable assembly of cores, fibers, and resins; it can be augmented with dimensions and tolerances, and other post-autoclave manufacturing information. This is very similar to how metallic parts are designed and manufactured. This approach has the additional benefit of separating the ply and component data that is potentially export controlled from the geometric form, fit, and function information needed for downstream applications that may involve a large supply chain.

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay.
"AN ANALYSIS OF RECENT STANDARDS ON COMPOSITE PRODUCT MODELS TO ENABLE DIGITAL TRANSFORMATION OF COMPOSITE PRODUCT MANUFACTURING."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

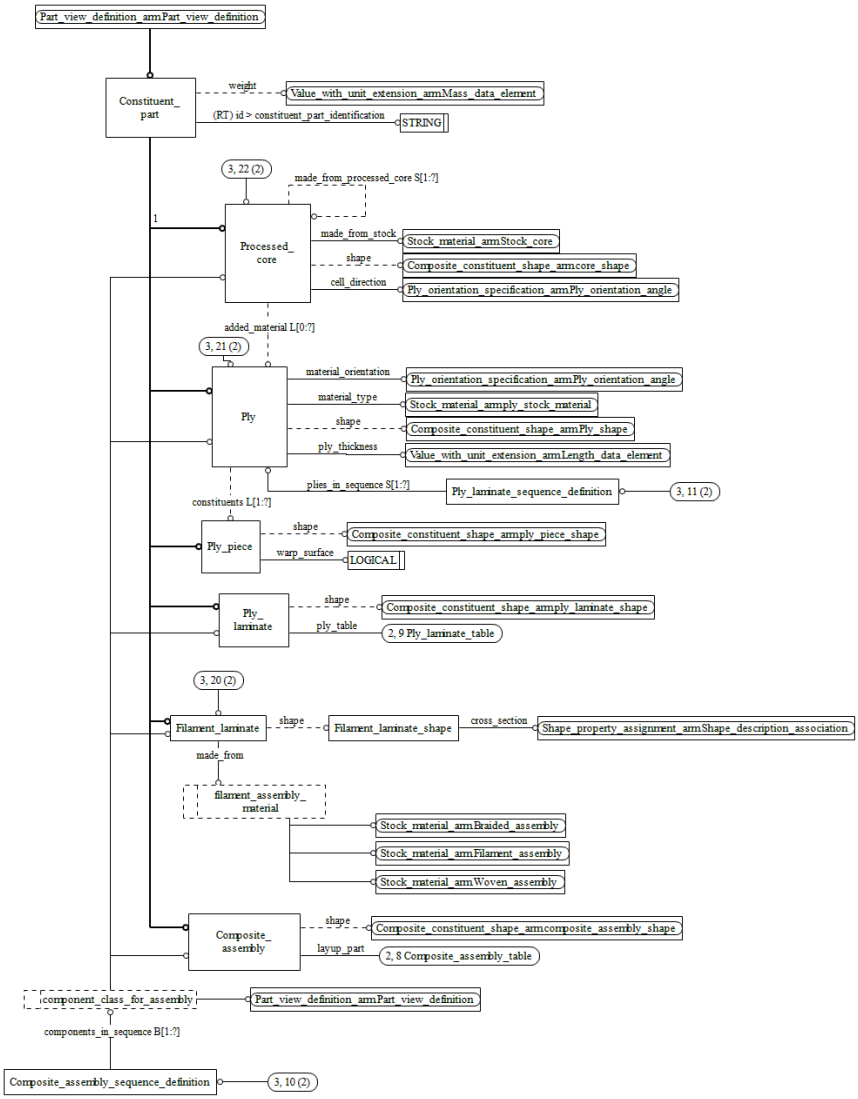


FIGURE 10: STEP data model for composite constituents [30, 31].

5. FUTURE RESEARCH AND STANDARDIZATION

The complexity and rapid technological changes in the FRP products and their production processes necessitate more research efforts and more revisions of standards. Even with the currently available standards, software vendor implementation and testing of the data models are needed to verify their compliance to the ASME and ISO standards. Such testing will involve both native models in the software vendor's proprietary system and the exchange models in the open ISO STEP formats.

Internal fiber arrangements and orientations of composite products are notoriously difficult to measure. In addition to destructive testing, several non-destructive testing methods are emerging to verify if the built part conforms to the design specifications. More research is needed for such measurements.

The new editions of ASME and ISO standards cover several composite manufacturing processes, but they need to cover even more processes that are being introduced into manufacturing at a rapid pace. There is an opportunity for both standards to be able to represent both a design (net) ply shape, and a manufacturing ply shape that included excess material added for activities such as hold-down pads for trimming and drilling. This calls for more research efforts in information modeling and engineering analysis tools.

6. SUMMARY AND CONCLUDING REMARKS

This paper presented a brief analysis of recent ASME and ISO standards on composite products. More details can be found in the standards documents themselves [7, 8]. By restricting the

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay.
"AN ANALYSIS OF RECENT STANDARDS ON COMPOSITE PRODUCT MODELS TO ENABLE DIGITAL TRANSFORMATION OF COMPOSITE PRODUCT MANUFACTURING."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

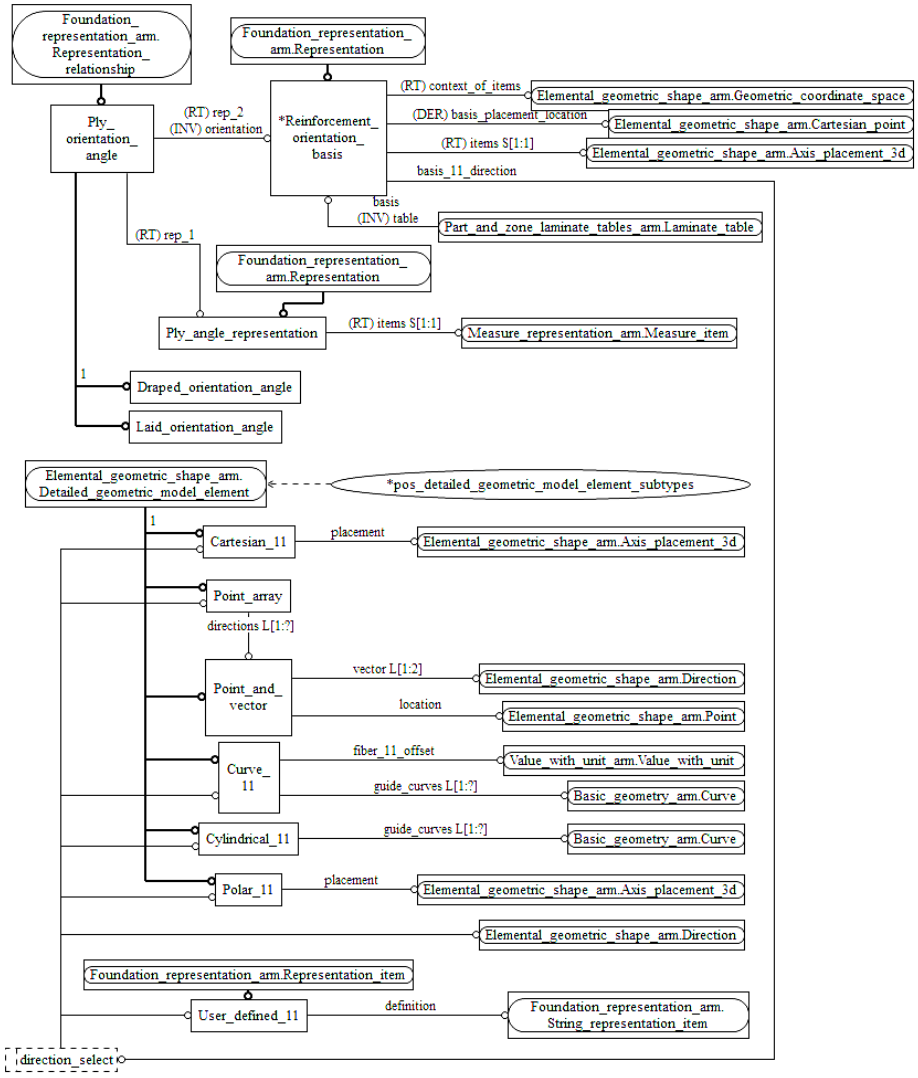


FIGURE 11: STEP data model for ply orientation [33].

analysis to a few salient features of these standards, the paper was able to focus on the concept of ply table, which together with its geometric companion, provides an important abstraction for information modeling.

While the ply table may represent only the uncured composite product, it provides a convenient representation for two of the important invariants under the curing process: combinatorial topology of the fibers and the angles between the fibers. Identification of these invariants in ply table is a major technical contribution of this paper. The ply table also provides vital information for sequencing for lay-up of plies, which is important for the composite manufacturing process.

The new editions of ASME and ISO standards are moving in the right direction towards enabling a digital transformation of composite product manufacturing. However, as described in Section 5, more research, measurements, and standardization are needed.

ACKNOWLEDGMENT AND A DISCLAIMER

The authors gratefully acknowledge the help and support of ASME and ISO standards committee members. Any mention of commercial products in this paper is for information only; it does not imply recommendation or endorsement by NIST.

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay.
"AN ANALYSIS OF RECENT STANDARDS ON COMPOSITE PRODUCT MODELS TO ENABLE DIGITAL TRANSFORMATION OF COMPOSITE PRODUCT MANUFACTURING."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

REFERENCES

[1] Ashby, M.F., 2011, *Material Selection in Mechanical Design*, 4th ed., Elsevier, Amsterdam.

[2] Ashby, M.F., 2009, *Materials and the Environment: Eco-Informed Material Choice*, Elsevier, Amsterdam.

[3] Helu, M., and Hedberg Jr, T., 2015. "Enabling smart manufacturing research and development using a product lifecycle test bed". Procedia Manufacturing, 1, pp. 86–97. DOI: 10.1016/j.promfg.2015.09.066

[4] Jeschke, S., Brecher, C., Song, H., and Rawat, D.B., (Editors), 2017, *Industrial Internet of Things – Cybermanufacturing Systems*, Springer, Switzerland.

[5] Thoben, K., Wiesner, S., and Wuest, T., 2017, "Industrie 4.0 and Smart Manufacturing – A Review of Research Issues and Application Examples," *Int. J. Automation Technol.*, Vol.11, No.1, pp. 4-16, DOI: 10.20965/ijat.2017.p0004

[6] Zhong, R. Y., Xu, X., Klotz, E., & Newman, S. T., 2017, Intelligent Manufacturing in the Context of Industry 4.0: A Review. *Engineering*, *3*(5), 616–630. DOI: 10.1016/J.ENG.2017.05.015

[7] ASME Y14.37-2019, *Product Definition for Composite Parts*, ASME, New York.

[8] ISO 10303-242:2019, Industrial automation systems and integration – Product data representation and exchange – Part 242: Application protocol: Managed model-based 3D engineering, International Organization for Standardization, Geneva, Switzerland.

[9] Campbell, F.C., 2010, *Structural Composite Materials*, ASM International, Materials Park, OH.

[10] Campbell, F.C., 2004, *Manufacturing Processes for Advanced Composites*, Elsevier, Oxford, U.K.

[11] ASME Y14.41-2019, Digital Product Definition Data Practices, ASME, New York.

[12] ISO 10303-42:2014, Industrial automation systems and integration – Product data representation and exchange – Part 42: Integrated generic resource Geometric and topological representation, International Organization for Standardization, Geneva, Switzerland.

[13] Wphallig. "Braids." Digital image. *Wikimedia Commons*. November 26, 2013. Accessed October 21, 2018. Licensed under CC BY-SA 3.0. https://commons.wikimedia.org/w/index.php?curid=29864409.

[14] Cross, N., 2009. "Gurit - a World Leader in Epoxy Prepreg Technology." Digital image. *Flickr*, January 30, 2009. Accessed October 25, 2018. Licensed under CC BY-ND 2.0. www.flickr.com/photos/80188450@N03/8138514559.

[15] Olson, G.B., 1997, Computational design of hierarchically structured materials, *Science*, Vol. 277, Issue 5330, pp. 1237-1242.

[16] ASME Y14.37-2012, *Composite Part Drawings*, ASME, New York.

[17] ASME Y14.5-2019, *Dimensioning and Tolerancing*, ASME, New York.

[18] SC4 On-Line. https://www.iso.org/committee/54158.html [accessed 10.23.2018].

[19] ISO 10303-1:1994, Industrial automation systems and integration—Product data representation and exchange—Part 1: Overview and fundamental principles. International Organization for Standardization, Geneva, Switzerland.

[20] Kemmerer S, editor., 1999, STEP the grand experience. NIST special publication, vol. 939. Washington, DC: US Government Printing Office. http://www.nist.gov/manuscript-publication-search.cfm?pub_id=821224 [accessed 10.23.2018].

[21] Srinivasan V., 2008, Standardizing the specification, verification, and exchange of product geometry: research, status and trends. *Computer Aided Des*, 40(7):738–49.

[22] Pratt MJ, Anderson BD, Ranger T., 2008, Towards the standardized exchange of parameterized feature-based CAD models, *Computer Aided Des*, 37(12):1251–65.

[23] Kim J, Pratt MJ, Iyer RG, Sriram RD, 2008, Standardized data exchange of CAD models with design intent. *Computer Aided Des*, 40(7):760–77.

[24] Kim BC, Mun D, Han S, Pratt MJ, 2011, A method to exchange procedurally represented 2D CAD model data using ISO 103030 STEP. *Computer Aided Des*, 43(12):1717–78.

[25] Wardhani R, Liu C, Mubarok K, Xu X., 2018, An Approach to Complete Product Definition Using STEP in Cloud Manufacturing. ASME. International Manufacturing Science and Engineering Conference, *Volume 1: Additive Manufacturing; Bio and Sustainable Manufacturing*, V001T05A021. doi:10.1115/MSEC2018-6613.

[26] Barnard Feeney, A., Frechette, S.P., and Srinivasan, V., 2015, A portrait of an ISO STEP tolerancing standard as an enabler of smart manufacturing systems. *Journal of Computing and Information Science in Engineering*, 15(2):021001–021001, ISSN 1530-9827. doi: 10.1115/1.4029050.

[27] Hunten, K.A., Barnard Feeney, A., Srinivasan, V., 2013, Recent advances in sharing standardized STEP composite structure design and manufacturing information, *Computer-Aided Design*, Vol. 45, pp. 1215–1221

[28] Barnard Feeney A., 2002, The STEP modular architecture. *ASME J Computer Inf Sci Eng*, 2(2):132–5.

[29] ISO/TS 10303-1770:2017-12(E), Industrial automation systems and integration — Product data representation and exchange — Part 1770: Application module: Part and zone laminate tables, International Organization for Standardization, Geneva, Switzerland.

[30] ISO/TS 10303-1767:2017-12(E), Industrial automation systems and integration — Product data representation and exchange — Part 1767: Application module: Composite constituent shape, International Organization for Standardization, Geneva, Switzerland.

[31] ISO/TS 10303-1768:2017-12(E), Industrial automation systems and integration — Product data representation and exchange — Part 1768: Application module: Composite material aspects, International Organization for Standardization, Geneva, Switzerland.

[32] ISO/TS 10303-1771:2014-02(E), Industrial automation systems and integration — Product data representation and exchange — Part 1771: Application module: Stock material, International Organization for Standardization, Geneva, Switzerland.

[33] ISO/TS 10303-1772:2017-12(E), Industrial automation systems and integration — Product data representation and exchange — Part 1772: Application module: Ply orientation specification, International Organization for Standardization, Geneva, Switzerland.

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay.
"AN ANALYSIS OF RECENT STANDARDS ON COMPOSITE PRODUCT MODELS TO ENABLE DIGITAL TRANSFORMATION OF COMPOSITE PRODUCT MANUFACTURING."
Paper presented at ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC 2019), Erie, PA, United States.
June 10, 2019 - June 14, 2019.

Proc. of the 10th Model-Based Enterprise Summit (MBE 2019), Gaithersburg, Maryland, USA, April 2-4, 2019

Securing, Authenticating, and Visualizing Data-Links for Manufacturing Enterprises

William Z. Bernstein
Systems Integration Division, NIST
Gaithersburg, MD, USA

Sylvere Krima
Engisis LLC
Bethesda, MD, USA

Laetitia Monnier
Universite de Bourgogne Franche-Comte
Dijon, France

Mehdi Shahid
TELECOM Nancy
Nancy, France

ABSTRACT

Managing digital resources generated from product design, manufacturing, and sustainment activities has become a significant burden for enterprises. In response, we introduce a prototype implementation of the Securing and Authenticating Data-Links (SADL) Interface, which interacts with a manufacturing handle registry to facilitate traceability of digital resources for engineering projects. This paper outlines the intended use of SADL and the handle registry by laying out hypothetical questions from potential users. Additionally, we map the core concepts of key standard data representations in manufacturing to a popular data type taxonomy. Future work will include the design and testing of data visualizations based on our mapping protocols.

INTRODUCTION

Manufacturing operations produce an immense amount of data, estimated at two exabytes of data annually in 2013 [1]. Considering the emergence of more complex electro-mechanical devices to the market, e.g., fully electric automobiles, the amount of code and manufacturing data is expected to grow significantly [2]. As a result, managing the associated digital resources has become, and will continue to become, a burden. An efficient and robust approach for labeling, categorizing, and curating diverse data is an essential first step for visualizing trends and deriving actions. In this paper, we present a prototype implementation of the manufacturing handle system, aimed at recording appropriate meta-data ("data about data") for digital resources related to the product lifecycle. Additionally, we present initial guidelines for developing data visualizations for the handle system to facilitate data exploration.

Each phase of the product's lifecycle incorporates its own data representations, organizational functions, and business processes. Each of these functions and processes uses tools and methods (e.g., computer-aided design (CAD) applications and requirement formalization). Though there have been efforts in improving information exchange across the various lifecycle phases (e.g., design and manufacturing) [3], there has not yet been a conclusive and robust demonstration at scale to achieve the so-called "digital thread" [4]. The metaphor "digital thread" conveys the seamless exchange and flow of data between engineering, manufacturing, business process and across supply chains [5].

However, in practice, the dearth of interoperability has led to gaps in information flow across manufacturing enterprises contributing to a number of challenges, including communicating across multi-tiered supply chains, reacting to engineering changes, and responding to customer requirements. For small-to-medium sized

DOI:https://doi.org/10.6028/NIST.AMS.100-24#page=59

enterprises, these challenges are even more difficult to overcome, since most solutions targeted at the Digital Thread are expensive, expert-driven one-off prototypes. In response to these challenges, we present the following research contributions: (1) a prototype interface, coined the Securing and Authenticating of Data-Links (SADL) Interface that allows users to register digital resources, add meta-data, and query the registry, (2) a classification scheme of lifecycle data from manufacturing phases, e.g., as-designed, as-planned, as-executed, and as-inspected, based on data nature and type [6], and (3) requirements on the further improvement of the SADL interface. It is our hope that this work facilitates a better digital resource certification management in a product's lifecycle for end-users, including plant managers, design teams, and supply chain managers.

Our efforts aim to facilitate the realization of a Model-Based Enterprise (MBE) that can quickly respond to product lifecycle disruptions, e.g., engineering change requests, weather events affecting suppliers, and machine degradation. From this perspective, we focus on the design, manufacturing, and inspection phases that incorporate a standard data representations, which have been primary focuses of the MBE journey. Managing these representations as digital resources, and changes to them, poses a significant challenge.

THE SADL INTERFACE AND THE MANUFACTURING HANDLE REGISTRY

SADL (Securing and Authenticating of Data-Links) is an application serving as a middleware between digital objects hosted on a Handle.Net registry and its end-users. Its goal is to offer a customizable overview of a product's lifecycle digital objects, the product data, by providing additional meta-data (e.g. lifecycle phase or product category) from which users can query, rank, order, classify, and construct links between objects.

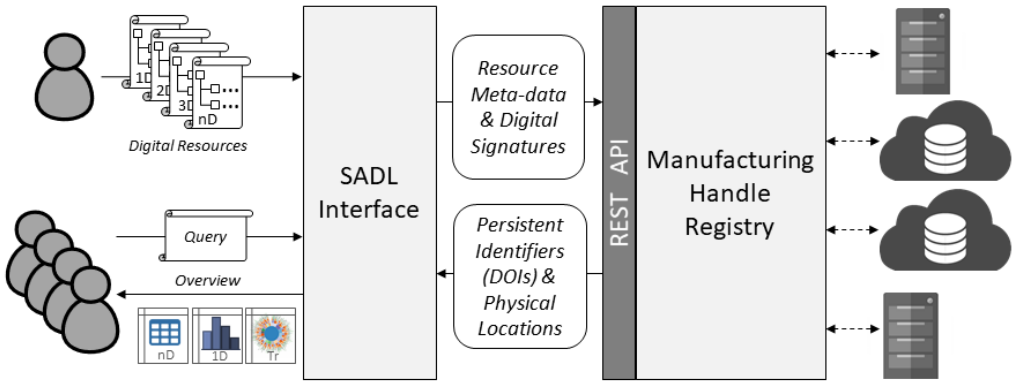


Figure 1: Vision of the SADL Interface and its interactions with users and the Handle Registry.

Figure 1 introduces the vision of the work presented in this paper. Initially, users label meta-data for digital objects by interacting with the SADL Interface. The SADL interface then leverages the Representational State Transfer (REST) Application Program Interface (API) of the Handle.Net registry to assign digital signatures and meta-data to the objects. The pipeline leverages existing technology that creates persistent identifiers (Digital Object Identifiers or DOIs [7]) to manage the physical locations of the resources. Other users can then submit queries through the SADL Interface to access report summaries of a collection of digital resources relevant to a manufacturing-oriented use case.

The pipeline presented above relies on the concept of the Handle System. Released by the Corporation for National Research Initiatives (CNRI) in 1994, the Handle System offers a means to locate, track, and manage data even in the face of constant modification [8]. In particular, manufacturing represents a domain that faces constant data modification and improvement. For instance, given engineering needs, it is very common that a

Bernstein, William; Krima, Sylvere; Monnier, Laetitia; Shahid, Mehdi.
"Securing, Authenticating, and Visualizing Data-Links for Manufacturing Enterprises."
Paper presented at 10th Model-Based Enterprise Summit (MBE 2019), Gaithersburg, MD, United States. April 2, 2019 - April 4, 2019.

product's design goes through several iterations before being approved for manufacturing, which leads to the creation of multiple design files. It is critical to construct a digital footprint outlining the complete history of each digital resource for various scenarios, e.g., liability investigations and engineering change requests. By providing a unique identifier for each digital object, the Handle System Registry acts as a DOI Repository providing a unique access point to all the digital resources inside an enterprise. Each "handle" (i.e., the name given to each of the DOIs) contains a set of meta-data about a specific digital object and allows for modifications to incur on the source file without compromising the validity and integrity of the handle.

By using the digital repository provided by the Handle System as a gateway to a broader view of all the different data objects inside the product's lifecycle, users can visualize the digital objects and link them together. This broader view describes the concept of a digital map encompassing critical digital resources given a particular use case. The Handle System provides digital record-keeping with a way to also offer efficient feedback to end-users and to facilitate better understanding of complex interactions with a product's lifecycle. Additionally, it is important that the current view in the registry reflects the current status in the real world. A proof-of-concept for a "System Lifecycle Handler" [9] confirmed the utility of these idea. In our pipeline, we include a means to store digital signatures to certify the validity of the stored data.

Throughout the rest of this paper, we explore potential means for visualizing a large collection of digital resources through the SADL Interface. We examine the expected range of data nature and type assuming that the registry is well-seeded with manufacturing-oriented data. This includes mapping expected business functions, from well accepted data representations to possible visualization schemes. We then conclude by addressing hypothetical, explorative questions from envisioned users to demonstrate the impact of the SADL Interface.

TOWARDS VISUALIZING A LARGE COLLECTION OF DIGITAL RESOURCES

In practice, the collection of digital resources for a given handle system would be quite large and a challenge to navigate. In response, we perform a data type mapping between the business processes and information embedded in standard data representations, namely STEP AP242 [10], STEP AP238 [11], MTConnect V1.4 [12], and QIF 3 [13]. STEP AP 242 provides an exchange format for design data including fully characterized Product and Manufacturing Information (PMI). STEP AP238 is a descriptive data representation for machine instructions, providing an additional layer of semantic descriptions compared to traditional G-code. MTConnect is a read-only communication protocol for capturing execution data from machine tool controllers. QIF is a semantically rich data format for representing, exchanging, and storing inspection plans, rules, and results. We applied the seven data type taxonomy used in Shneiderman's task-by-type taxonomy [6] keeping in mind that future work will involve designing and implementing data visualizations to respond to user queries to the SADL Interface. Nomenclature and a brief description of all seven data types are below.

***1-dimensional* (1D):** linear data types that are organized in a sequential manner, e.g., textual documents that only contain alphabetically ordered strings. From the perspective of engineering design, 1D data types can relate to the rules and requirements needed to apply STEP AP242. Considering the rules and requirements elements, a Mural visualization in the background of the scrollbar [14] might be appropriate. This visualization would highlight different parts of the text that are related to a particular requirement.

***2-dimensional* (2D)**: planar or map data including maps, floorplans, or newspaper layouts. Manufacturing-oriented examples could include entire factory layouts or plans for individual cell layouts. In a robotic factory, a production cell layout could highlight the distance between the robotic arm and other assets. A higher-level

Bernstein, William; Krima, Sylvere; Monnier, Laetitia; Shahid, Mehdi.
"Securing, Authenticating, and Visualizing Data-Links for Manufacturing Enterprises."
Paper presented at 10th Model-Based Enterprise Summit (MBE 2019), Gaithersburg, MD, United States. April 2, 2019 - April 4, 2019.

instance of 2D data could represent the entire plant factory decomposed into multiple cellular maps each relating to one another.

***3-dimensional* (3D)**: real-world objects or representations that represent volumetric data, such as solid models or computer-aided design (CAD) files. From a visualization perspective, the challenge is to find a balance between the view of the real-life object and the information within it. The aspect of the object must correlated with the data in a way that the user understands its spatial positioning in a larger context, e.g., a complex assembly.

***Multi-dimensional* (nD):** relational and statistical databases manipulated as multidimensional data in which items with *n* attributes become points in a *n*-dimensional space. QIF inspection results, including probe information are an example. One visualization technique to represent a database table is a cube visualization [15]. Each face of the cube is composed of one attribute of the table and layers of each face correspond to a possible value of the face attribute. The intersection of two adjacent faces represents a data point.

Temporal* (Ts)*: time series data, such as historical presentations or future projections. The difference between 1D and temporal data is rather nuanced. Both can be simple text documents, but the main difference is that Ts data is anchored through a timeline. Process plans in STEP AP238 and sample type data in MTConnect streams are examples of Ts data types.

Tree* (Tr)*: hierarchies with each item having a link to one parent item except the root. For example, an MTConnect Device model is organized based on the design of devices. Using a tree visualization, like a Tree-map [16], could provide a snapshot of capabilities and characteristics of available devices.

Network* (Nk)*: items linked to an arbitrary number of other items not following rules of trees. For example, the STEP AP242 Assembly structure contains parts and subassemblies as items, and multiple type of links can exist between these items. Some links might have numerical values or represent specific actions or modifications. A matrix diagram [17] is a way to expose different items and their associated links. The matrix can be color coded so that the user has a better understanding of the differences in type and meaning of the links.

			Data Types						
Representation	**Business Function**	**Concept Description**	**1D**	**2D**	**3D**	**nD**	**Ts**	**Tr**	**Nk**
STEP 242 (as-designed)	Specification, Breakdown & Configuration	Assembly Structure						●	●
		Transformations, Geometry, & Coordinate System		●	●				
	… … …								
STEP 238 (as-planned)	Model-Based Manufacturing Process	Generic Toolpaths		●	●				
		Parameters (feeds, speeds, etc.)	●						
	… … …								
MTConnect (as-executed)	Historical Machine Operations	Samples					●		
		Conditions	●						
	… … …								
QIF (as-inspected)	Model-Based Definition	Computer-aided design (CAD) data			●			●	
		Product manufacturing information (PMI) data				●			
	… … …								

Figure 2: Classification of data type per business function and concept description within each studied data representation. The complete table[1] can be accessed here: https://goo.gl/Zbkqmb.

[1] An initial draft of the mapping. We expect it to evolve as we dive deeper into each data representation.

Bernstein, William; Krima, Sylvere; Monnier, Laetitia; Shahid, Mehdi.
"Securing, Authenticating, and Visualizing Data-Links for Manufacturing Enterprises."
Paper presented at 10th Model-Based Enterprise Summit (MBE 2019), Gaithersburg, MD, United States. April 2, 2019 - April 4, 2019.

Figure 2 illustrates the structure of the data type classification completed for each business function of studied standard representations. We conducted this classification to identify design opportunities for custom visualizations for the SADL Interface. Besides individual visualizations per business function, we also envision potential for effective overview visualizations. In other words, given that an organization registered a large amount of digital resources within the handle system, we can present, for example, a hierarchical representation based on a prominent concept description, e.g., the assembly structure of a product. In the future, we plan to implement and test the effectiveness of sunburst plots, cartesian node-link diagrams, and matrix views [18] using accepted information visualization principles [19].

FROM DATA SETS TO VISUAL INDICATORS

In the previous section, we introduced a mechanism to uniquely identify, locate, authenticate, and navigate through a product's lifecycle digital objects using the Handle.Net and our SADL interface. We also described different data visualization types. In this section, we will discuss some steps towards generating and visualizing insight and performance metrics from trusted product data. While our area of focus is limited to as-designed, as-planned, as-executed, and as-inspected lifecycle data, the ideas and methods introduced here can be applied to other data.

The first step consists of identifying and categorizing the different product data generated and available to the organization. During this step, one must ensure that (1) the data is complete and fits under one of the four lifecycle stages previously mentioned, (2) the data is available in an open-standard format to reduce the cost of processing and enable interoperability, and (3) the data concepts are properly identified (as in the third column in Fig. 2).

The second step consists of identifying metrics or indicators based on key organizational characteristics and derived from the data concepts previously identified. These metrics and indicators should provide answers to questions related to organizational resources, activities, and performance. In this project, we have focused our questions on basic organizational components, namely processes, products and people (or the 3Ps). The following is a set of hypothetical questions illustrating some basic metrics and indicators that can easily be computed and/or inferred from the data sources/standards we use:

1. **Process**:
   1.1. How long did it take to execute process X during the past 10 days?
   1.2. How many parts a day are handled during process X?
   1.3. Was there a quality improvement between V2 and V1 of process X?
2. **Product**:
   2.1. What was the assembly structure of Product Y?
   2.2. How many parts were affected after changing feature X on product Y?
   2.3. Was the new design of Product X actually ready to move to production on November 2, 2018?
3. **People**:
   3.1. Who inspected the version of part Z that was built on November 2, 2018?
   3.2. What was the chain-of-command for Product X through its lifecycle?

The third and last step maps the different questions from the second step to the right source of data, data concepts, and visual data types to address each question. The output should be similar to Table 1 and be used as a guideline for the solution implementer(s). Our recommended output is a Table with the following columns:

Bernstein, William; Krima, Sylvere; Monnier, Laetitia; Shahid, Mehdi.
"Securing, Authenticating, and Visualizing Data-Links for Manufacturing Enterprises."
Paper presented at 10th Model-Based Enterprise Summit (MBE 2019), Gaithersburg, MD, United States. April 2, 2019 - April 4, 2019.

1. **Question**: the question whose answer is a metric or indicator regarding a key organizational component
2. **Representation** or data source: the format of the data that will be used to compute the metric or indicator in response to the question
3. **Key concepts**: a list of the data elements that need to be extracted from the data source to compute the metric or indicator
4. **Shneiderman's Data Type**: the data type used to visualize the metric or indicator based on the list of types identified in the previous section

In this section, we presented a three-step process to guide a user from identifying the type of data sources available, derive performance metrics and indicators, and map them to a visual data type using the Shneiderman's classification. This process would facilitate the design and development of a visualization dashboard providing insight to engineering teams through open data representations.

Table 1: Recommended output to describe requirements for presenting relevant data through the SADL Interface based on an enterprise-driven question.

Question	Representation(s)	Key concept(s)	Shneiderman's Data Type
1.1	MTConnect	Events	Ts
1.2	MTConnect	Part Count, Samples	1D
1.3	MTConnect, QIF	Events, Measurements data	1D
2.1	AP242	Assembly structure	Tr
2.2	AP242	Assembly structure	Tree/1D
2.3	AP242	Meta-data entered at SADL*	Tr
3.1	AP242, QIF	General Mgmt. Information	1D
3.2	AP242	Meta-data entered at SADL*	Ts/Tr

*Not part of the standard representation itself. The digital signatures would be appended to the digital resource once the user enters the information in the SADL Interface.

CONCLUSIONS

We presented progress towards the SADL, an interface designed to secure and authenticate data-links within a standardized handle system. In doing so, we stressed the importance of implementing effective and interactive visualizations that aid users in querying a large collection of digital resources. We demonstrate this process across four leading standards for the model-based enterprise: STEP AP242, STEP AP238, MTConnect, and QIF. We expect that others can follow the same process for other standards as well. Relating the underlying domain-specific data to a taxonomy of domain-agnostic data types eases the integration of state-of-the-art visualizations. Such visualizations are expected to be integrated within the SADL interface. Future work will consider the feasibility of generalized visualizations so that a variety of digital resources can be represented to enhance organizational decision-making.

Bernstein, William; Krima, Sylvere; Monnier, Laetitia; Shahid, Mehdi.
"Securing, Authenticating, and Visualizing Data-Links for Manufacturing Enterprises."
Paper presented at 10th Model-Based Enterprise Summit (MBE 2019), Gaithersburg, MD, United States. April 2, 2019 - April 4, 2019.

DISCLAIMER

No endorsement of any commercial product by NIST is intended. Commercial materials are identified in this report to facilitate better understanding. Such identification does not imply endorsement by NIST nor does it imply the materials identified are necessarily the best available for the purpose.

ACKNOWLEDGEMENTS

We thank Moneer Helu, Robert Lipman, and Tom Kramer for their valuable feedback that improved the paper.

REFERENCES

[1] Auschitzky, E., Hammer, M., & Rajagopaul, A. (2014). How big data can improve manufacturing. *McKinsey & Company*, *822*.

[2] Yin, S., & Kaynak, O. (2015). Big data for modern industry: challenges and trends [point of view]. *Proceedings of the IEEE*, *103*(2), 143-146.

[3] Panetto, H., & Molina, A. (2008). Enterprise integration and interoperability in manufacturing systems: Trends and issues. *Computers in industry*, *59*(7), 641-646.

[4] West, T. D., & Blackburn, M. (2017). Is Digital Thread/Digital Twin Affordable? A Systemic Assessment of the Cost of DoD's Latest Manhattan Project. *Procedia Computer Science*, *114*, 47-56.

[5] Hedberg, T., Lubell, J., Fischer, L., Maggiano, L., & Feeney, A. B. (2016). Testing the digital thread in support of model-based manufacturing and inspection. *Journal of computing and information science in engineering*, *16*(2), 021001.

[6] Shneiderman, Ben. "The eyes have it: A task by data type taxonomy for information visualizations." *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996.

[7] Paskin, N. (2010). Digital object identifier (DOI®) system. *Encyclopedia of library and information sciences*, *3*, 1586-1592.

[8] Kahn, Robert, & Wilensky, Robert, "A framework for distributed digital object services," May 13, 1995. URL: http://www.cnri.reston.va.us/home/cstr/arch/k-w.html. Accessed December 1, 2018.

[9] Bajaj, Manas, and Thomas Hedberg Jr. "System Lifecycle Handler—Spinning a Digital Thread for Manufacturing." *INCOSE International Symposium*. Vol. 28. No. 1. 2018.

[10] ISO (2014). 10303-242: 2014, Industrial automation systems and integration—product data representation and exchange—Part 242: Application protocol: Managed model based 3d engineering. *Geneva (Switzerland): International Organization for Standardization (ISO)*.

[11] ISO (2007). 10303-238: 2007, Industrial automation systems and integration—product data representation and exchange—Part 238: Application protocol: Application interpreted model for computerized numerical controllers. *Geneva: International Organization for Standardization (ISO)*.

[12] Sobel, W. (2015). MTConnect standard. Part 1—overview and protocol. *Standard—MTConnect.* URL: http://www. mtconnect. org/standard*.* Accessed December 1, 2018. .

[13] Dimensional Metrology Standards Consortium. (2018). Part 1: Overview and Fundamental Principles in Quality Information Framework (QIF)—An Integrated Model for Manufacturing Quality Information. *Dimensional Metrology Standards Consortium*. URL: http://qifstandards.org/. Accessed Dec 1, 2018.

[14] Jerding, Dean F., and John T. Stasko. "The information mural: A technique for displaying and navigating large information spaces." *IEEE Transactions on Visualization and Computer Graphics* 4, no. 3 (1998): 257-271.

[15] Stolte, Chris, Diane Tang, and Pat Hanrahan. "Multiscale visualization using data cubes." *IEEE Transactions on Visualization and Computer Graphics* 9.2 (2003): 176-187.

[16] Shneiderman, Ben. "Tree visualization with tree-maps: 2-d space-filling approach." *ACM Transactions on graphics (TOG)*11.1 (1992): 92-99.

[17] Van Ham, Frank. "Using multilevel call matrices in large software projects." *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*. IEEE, 2003.

[18] Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Commun. Acm*, *53*(6), 59-67.

[19] Carpendale, M. S. T. (2003). Considering visual variables as a basis for information visualisation.

Platform-Independent Debugging of Physical Interaction and Signal Flow Models

Mehdi Dadfarnia
Engineering Laboratory
National Intitute of Standards & Technology
Gaithersburg, USA
mehdi.dadfarnia@nist.gov

Raphael Barbau
Engisis, LLC
Bethesda, USA
raphael.barbau@nist.gov

***Abstract*—**Systems engineering tools are used to organize development activities of a wide variety of engineers, many of whom develop discipline-specific simulation models. To increase the efficiency of this process, systems modeling tools have been extended to represent physical interactions and signal flows that can be translated to simulation tools and executed. Sometimes these simulation models fail to execute or they produce unexpected execution results. It is helpful to identify causes of these problems in earlier stages of system model development, before they propagate to fully-developed simulation models. Debugging physical interaction models is difficult because their execution is bidirectional between system components. This paper gives an overview of debugging procedures for physical interactions and (unidirectional) signal flows in platform-independent system models that integrate with domain-specific simulation models. These procedures identify system model causes of simulation execution failure or incorrect simulation results.**

Keywords—SysML, debugging, modeling, simulation, lumped parameter, equation-based languages

I. Introduction

The increasing complexity of modern engineered systems and products requires integrating systems modeling and simulation tools to improve efficiency of design processes [1]. These tools capture behavioral and structural aspects of systems or products that are checked (analyzed) without prototyping and experimenting on real systems [2]. Systems engineers use systems modeling tools to organize and coordinate analysis by a wide variety of engineers [3], many of whom develop their own equation-based simulation models [4].

Some simulations tools present graphical interfaces showing interconnection of components, corresponding to energy and information exchange between physical objects in the real system [4, 5]. They are referred to as *physical interaction and signal flow* models (also known as lumped parameter, one-dimensional, or network models) [6]. These models help manage system complexity by focusing on *what* systems should accomplish, rather than *how* [5]. Simulations of these models are experiments that answer questions about the systems being modeled without physically building them [4, 5].

Many simulation tools incorporate equation-based modeling languages for physical interaction and signal flow [7, 8, 9, 10]. System component specifications include ordinary and algebraic differential equations, while their interconnections generate additional equations between components. These models can represent a wide range of discipline-specific physical interactions between components (mechanical, electrical, etc.) as well as communication of numeric and Boolean information.

Systems modeling tools also focus on how components are interconnected and broken down into subcomponents (system structure) [11]. Interconnections between components reflect physical interactions and information exchanges. This organization enables systems engineers to coordinate others in specialized engineering disciplines, focusing on subsets of components and interconnections that require their expertise [12]. However, unlike simulation tools, system models are not strictly equation-based.

System models are often developed separately from simulation, leading to inconsistent specifications of overlapping aspects of the system [6]. To minimize these inconsistencies and enhance model interoperability, we developed a publicly-available extension to the Systems Modeling Language (SysML) [13] (the most widely used graphical modeling language for systems engineering) that facilitates platform-independent integration of SysML with physical interaction and signal flow simulation tools (SysML Extension for Physical Interaction and Signal Flow Simulation, SysPhS) [6, 14]. We developed software to translate these extended SysML models into simulation files that run on two widely-used simulation platforms [15].

Despite the increased ease of use, efficiency, and integration provided by the extension and translator, it is often hard to identify (debug) the cause of errors in the models. Determining the cause of errors is a critical step in correcting models. This paper is concerned with identifying causes of failure to:

- Execute simulation models translated from system models,
- Get expected results from simulation execution.

The debugging procedures presented[1] focus on system models written in SysML that are translated to simulation, rather than on particular kinds of simulation models. The procedures are independent of simulation platform (language or tool), aiming to fix errors in system models before they spread to simulation models on multiple platforms. Failures of translation from system models to simulation due to incorrect usage of the SysPhS extension or errors in the translator are not considered.

Integrating SysML models of physical interactions and signal flows with simulation tools enables compilation, simulation, and validation of the SysML model. System models can be checked for failure to compile and simulate a translated model or failure to generate expected results from

The work of Raphael Barbau was suppo ted by U.S. National Institute of Standards and Technology grant award 70NANB14H249 to Engisis, LLC.

[1] Author Raphael Barbau developed these techniques.

Dadfarnia, Mehdi; Barbau, Raphael.
"Platform-Independent Debugging of Physical Interaction and Signal Flow Models."
Paper presented at The 13th Annual IEEE International Systems Conference, Orlando, FL, United States. April 8, 2019 - April 11, 2019.

the simulation run-time. Although the causes of these two kinds of errors might be similar, debugging techniques in system models differs from those used in equation-based modeling languages.

Section II categorizes errors and debugging techniques for each kind of error. It also surveys literature on debugging techniques for system models and simulation tools for equation-based modeling. It describes SysML and SysPhS features that are important to debugging physical interactions and signal flows, using examples from a cruise control system. Section III gives an overview of the proposed debugging procedures. Section IV concludes with a discussion of our findings, and an outlook for future work.

II. Background

A. Types of Errors and Debugging Techniques in Physical Interaction and Signal Flow Models

Several types of errors can cause failures in physical interaction and signal flow models in systems and equation-based modeling languages. This paper focuses on identifying errors that cause simulation to fail or generate unexpected results during execution. The type of failure influences the debugging procedures required.

Errors that cause failure to simulate arise from model structure. These show the modeler's design does not properly support simulation. The underlying system of equations is inconsistent, including overconstrained (more equations than variables) or underconstrained (fewer equations than variables). It also includes equations that would divide by zero, functions being called outside of their real domain (such as the square root of a negative number), and other erroneous symbolic transformations.

Errors that cause the simulation to produce unintended results arise from the meaning of the model. These reflect discrepancies between desired system behavior and simulation execution. Although some errors can be identified automatically (such as variable values outside bounds) depending on the simulation tool being used, these errors can also be found manually after trying to validate the simulation results. These errors can come from incorrect equations, incorrect parameter or initialization values, and incorrect function calls from equations. Errors can also be due to integration errors in the equation solvers being used [1], which will not be discussed in this paper.

The difficulty in debugging these errors in physical interactions is that observing ordered execution of command sequences or operations do not work in models that include bidirectional relationships. This fundamental difference means that debuggers for errors in physical interactions need to examine chains of variable transformations in the model.

Static debugging techniques identify errors that cause failure to compile simulation models to executable code. These techniques trace symbolic transformations through the model to identify erroneous sections [1]. Dynamic debugging techniques identify errors that cause simulation to produce unexpected results. These techniques involve interactive inspection of models during execution [1]. They must be used after static debugging techniques ensure the model can be compiled to executable code.

B. State-of-the-art Debugging Techniques for Physical Interaction and Signal Flow Models

Though methods for identifying errors in system specifications share many similarities [5, 16], models that include physical interaction require specialized debugging strategies because these interactions affect all components involved. This is in contrast to sequential execution of operations in most other modeling languages, including signal flow simulation, where each component only affects the ones executing after it. Since models of physical interaction always have two participants we call it bidirectional (sometimes known as "acausal" even though causality applies to all physics), while models of signal flow are unidirectional.

Debugging in bidirectional and unidirectional modeling requires different approaches [5]. In unidirectional models, an error found at some point in the execution implies that the cause of the error is in one of the components or operations executed prior to that point. Errors in bidirectional models do not come from a past sequence of executed operations or components, because they are not executed in sequence.

Identifying errors in bidirectional (physical interaction) models has received significant attention in the Modelica language community [17]. In [1, 18, 19, 20, 21], authors look at debugging models in equation-based languages, as well as scalability of such techniques for larger models. Earlier work focused on code instrumentation to provide traditional debugging mechanisms such as breakpoints and single-stepping [20], which are more useful for sequential languages. There is also work focused on determining whether a system of equations is balanced [18], as well as techniques to semi-automatically isolate data flow slices to find potential sources of failure [19].

In [1, 22, 23], authors integrate information from variable (symbolic) transformations from static debugging into a dynamic debugger at simulation run-time (transformations are mathematical operations on variables to give values to other variables). The debuggers have graphical interfaces for exploring variable simulation and traditional debugging techniques such as breakpoints and single-stepping. Techniques for tracing symbolic transformation are critical to debugging physical interaction and signal flow models, including those translated from system models.

Implementing these debugging techniques in system models is more difficult than in simulation because of the higher level of abstraction preferred in early stages of systems engineering processes. Unlike equation-based modeling languages, system models specify multi-disciplinary, conceptual models of a system and its components. Tracing symbolic transformations requires a more diagram-based procedure in these models. The authors in [24] describe the structure of a functional, system model debugger that integrates system models with an equation-based modeling language through a mapping between them. The debugger focuses on visualizing variables running through a simulation, and much less on symbolic transformations. This paper argues that some tracing of symbolic transformations is necessary in system models.

It is burdensome to coordinate changes between system models and equation-based simulation models. One way to coordinate changes is to fix errors in simulation tools and feed the corrections back into system models. Another way is

Dadfarnia, Mehdi; Barbau, Raphael.
"Platform-Independent Debugging of Physical Interaction and Signal Flow Models."
Paper presented at The 13th Annual IEEE International Systems Conference, Orlando, FL, United States. April 8, 2019 - April 11, 2019.

to debug system models before generating and experimenting with simulation models. This paper focuses on the latter approach. Although both are useful, applying equation-based model debugging techniques to system models at earlier development stages can verify and increase understanding of the relationships captured in system models before discipline-specific experts focus on parts of the systems in their own models and tools. It also helps fix errors in functional, higher abstraction, platform-independent system models before they spread to behavioral, lower abstraction, domain-specific simulation models.

Next, we give some background information about modeling physical interactions and signal flows in SysML that is essential for tracing symbolic transformations. In the subsequent section, we describe the the debugging techniques applied to SysML system models.

C. *Modeling Physical Interactions and Signal Flows with SysPhS*

The SysML extension and simulation translator we developed [14, 15] integrates SysML with physical interaction and signal flow modeling simulation by identifying modeling capabilities in common between simulation platforms, comparing those with SysML, and extending SysML with only those modeling capabilities that SysML does not have [6]. Development starts with system models in extended SysML, then translates them into simulation platforms. This means that any errors not due to usage of the extension, translator, or simulator's execution engine will be in the SysML model.

Static debugging requires tracing through variable transformations in the model, and dynamic debugging uses bookkeeping of values over simulated time to pinpoint errors. Tracing physical interactions and signal flows in SysML system models extended with SysPhS [14] is done through connectors. The role of connectors in modeling physical interactions and signal flows is briefly discussed in this section; more can be found in [6, 12, 14].

Modeling in SysML starts with system components and their interactions in an internal block diagram (IBD), as in Figure 1. It shows physical interaction and signal flows involved in automatic control of a vehicle's speed, between the vehicle, its components, the cruise controller, and the operating environment. Interactions are represented by SysML connectors, which show exchanges of physical substances or signals occurring between the ends of each connector. The ends are either parts, or ports (smaller rectangles appearing on rectangles) that are placed on parts, on connector properties, or on other ports. Part names appear in titles of the rectangles in IBDs, before the colons. Each part is a role in the model (such as *driver* in Figure 1) and is played (typed) by a kind of thing (such as *Person*), appearing in the title after a colon and represented by a SysML block. Ports are essentially parts of parts, playing roles of roles in a model. They are also typed by blocks to show the kind of thing playing each role.

Item flows on connectors are optional, but useful to represent the type of signal or conserved physical substance flowing between parts or ports. They appear in Figure 1 as labels of filled triangles on connector lines, such as the *LinearMomentum* label between the parts *gravVehicleLink* and *controlledVehicle*. One filled triangle on a connector indicates signal flow in the direction of the triangle is pointing, while two pointing in opposite directions indicate physical interaction.

Blocks that type parts or ports at the ends of connectors must have flow properties typed by the kind of signals or physical substances flowing. Flow properties or

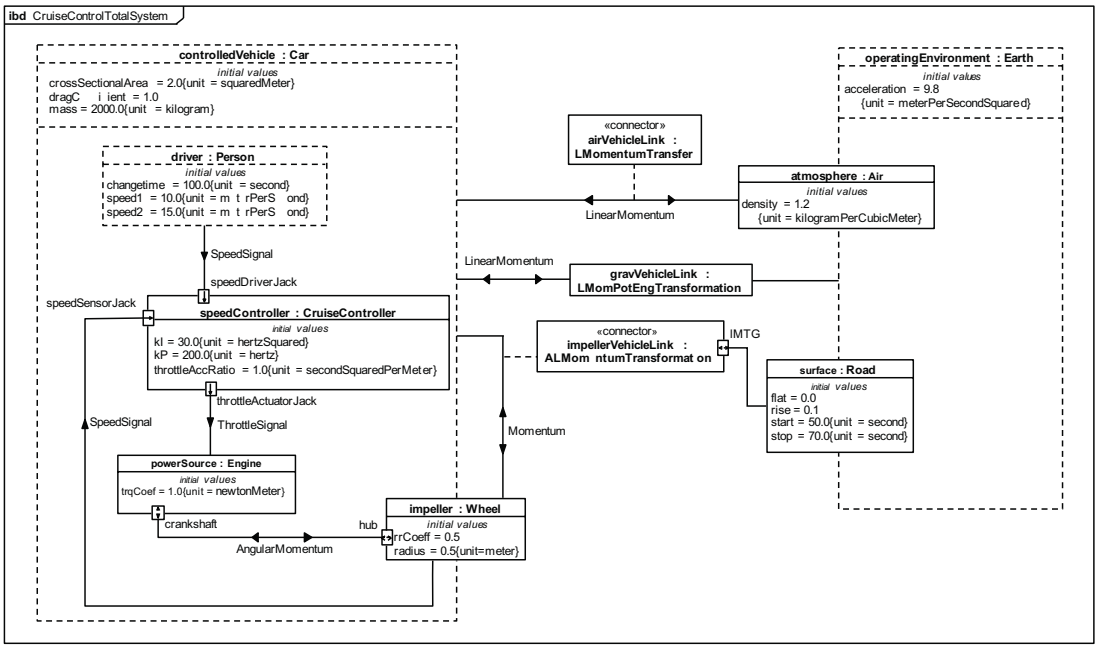


Fig. 1. Internal Block Diagram for a vehicle cruise control system, defined in SysML extended by SysPhS for physical interaction and signal flow

Dadfarnia, Mehdi; Barbau, Raphael.
"Platform-Independent Debugging of Physical Interaction and Signal Flow Models."
Paper presented at The 13th Annual IEEE International Systems Conference, Orlando, FL, United States. April 8, 2019 - April 11, 2019.

their types represent variables (either conserved or non-conserved, see below). Flow properties for signal flows are *in* or *out* (unidirectional) variables, typed by the kind of signal flowing. Flow properties for physical interactions are *inout* (bidirectional), typed by the physical substances flowing. These substances have variables for the substance's flow rate and potential to flow (one variable for each). Blocks that type parts or ports at the ends of connectors can be selected from libraries of reusable component interaction blocks or be specified by modelers. They can include more properties besides the flow properties, as well as multiple flow properties if the part or port typed by that block is involved with multiple types of signal flows or physical interactions.

The SysPhS extension includes a *PhSConstant* stereotype to specify that property values are to remain constant during each simulation execution. It also includes a *PhSVariable* stereotype to specify that they might vary during simulation. Flow properties for physical interaction are typed by blocks from a SysPhS library that have properties with *PhSVariable* stereotypes applied. Flow properties for signal flows have *PhSVariable* stereotypes applied as well. More details on defining blocks and their properties can be found in the SysPhS specification [14].

Connectors also imply a mathematical relationship between variables of flow properties defined by the blocks typing parts or ports at each end of a connector. After connectors are translated to simulation, those tools generate equations for them (differently for physical interaction and signal flow, see below). Parts and ports may also perform mathematical manipulations on flow property variables, specified by the blocks that type them, using SysML's parametric diagrams. These diagrams equate (bind) variables of their equations to flow properties and other properties from their containing blocks. Equations in these diagrams are called constraints and their variables are called constraint parameters.

In physical interactions, when connected ports and parts are typed by blocks that have the same *inout* flow property (and no connector properties are involved, see below), the values of conserved variables (flow rates of conserved substances) for the same flow property must add to zero during simulation, while their paired non-conserved variables (potential to flow) must be the same. Connectors can be augmented by a property of the containing block that represents physical substances transferred between parts or ports in the system and not at the boundary of any object. A connector property can also represent transformation of one type of physical substance – and equivalently one set of flow property variables – to another. The transfers and transformations represented by connector properties are specified with equations in parametric diagrams. More about connector properties can be found in [6, 12, 14].

In signal flows, when connected parts and ports are typed by blocks with the same flow property but opposite directions (one *in*, the other *out*), the flow properties (which are non-conserved variables in this case) must have the same value during simulation. Multiple connectors cannot have the same *in*-flow property at their ends because signals flowing into it would conflict.

Lastly, connectors linking to a port or part with no flow properties, such as the one between parts *gravVehicleLink* and *operatingEnvironment* in Figure 1, are structural relationships that enable physical interaction, but across which not physical substances flow. In Figure 1 these are the connectors directly to the earth and road, which are necessary for the car to store/consume potential energy and to propel itself, respectively, but are assuming to be immovable themselves.

III. Debugging Methods

The SysPhS extension enables translation from system models to equation-based models. If an equation-based model fails to compile or simulate correctly, the cause can be identified by tracing through chains of connectors between components. This is the basis for static debugging techniques and facilitates dynamic debugging. Before overviewing the techniques, we discuss a procedure to simplify system models by distinguishing between physical interaction and signal flow connectors. Simplifying models beforehand makes debugging more straightforward and scalable.

A. Preprocessing: Simplifying Models

The system model is broken down into one for physical interactions and another for signal flows. This enables separate debugging of two simpler system models before replicating the resulting fixes in the complete model, a simpler task than debugging the entire model all at once.

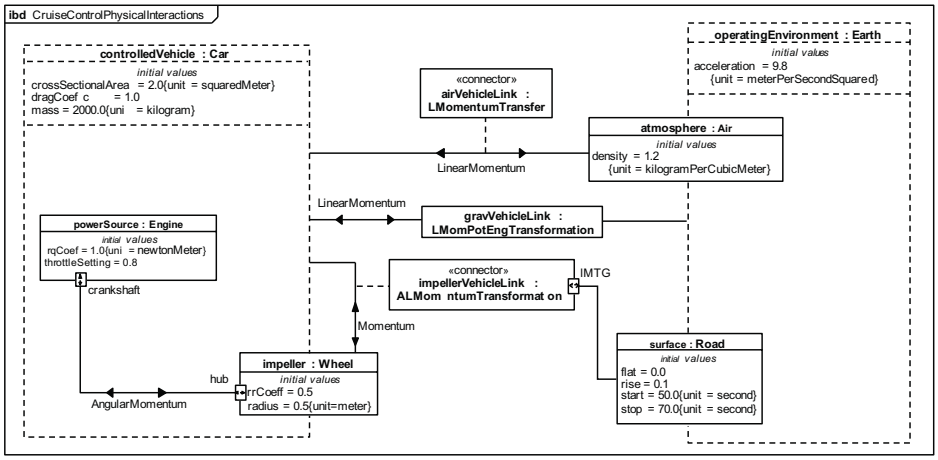


Fig. 2. Cruise control model with only physical interactions

Dadfarnia, Mehdi; Barbau, Raphael.
"Platform-Independent Debugging of Physical Interaction and Signal Flow Models."
Paper presented at The 13th Annual IEEE International Systems Conference, Orlando, FL, United States. April 8, 2019 - April 11, 2019.

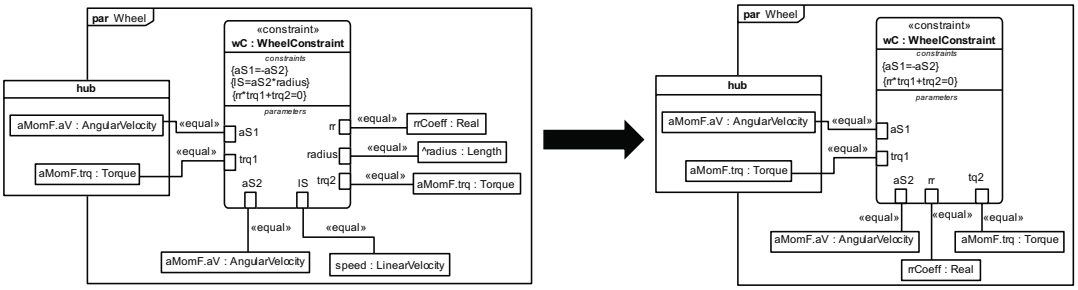


Fig. 3. Show two (2) parametric diagrams of the same system (before and after changes)

First, a model of physical interactions in the system is derived by eliminating all connectors in the original model's IBDs that do not represent physical interactions (saving a separate copy of the original model first). Any remaining parts or ports that are not at the end of the remaining connectors or do not possess a port that is at the end of a remaining connector are also eliminated. Following the example in Figure 1, Figure 2 presents an IBD with only the physical interactions in the original cruise control system.

Next, in each parametric diagram for the remaining parts or ports, eliminate equations (constraints) determining values of variables (constraint parameters) that are bound to (signal flow) *out*-flow properties. Eliminate part or port properties that are bound to variables on these *out*-flow equations as well. Replace any remaining equation variables bound to *in*-flow properties on the parts or ports by constant values, either by directly replacing the parameter with a constant value in constraints or by introducing a binding to a *PhSConstant*-stereotyped property that has a constant default value or instance value (see [14] for value assignment examples). Figure 3 depicts a parametric diagram for a component in Figure 2, before and after these changes were made.

A separate system model for signal flows is derived by first eliminating all connectors in the original model's IBDs that do not represent signal flows (while saving a separate copy of the original model). Also eliminate any remaining parts or ports that are not at the end of the remaining connectors or does not possess a port that is at the end of a connector.

Next, in each parametric diagram for the remaining parts or ports, eliminate equations that play no role in determining values of variables bound to *out*-flow properties or equations that do not have any bindings to *in*-flow properties are eliminated. Part or port properties not bound to variables on the remaining equations. Of the remaining equations, some variables might be bound to physical interaction *inout*-flow properties on the parts or ports. These flow properties are replaced in simplification. If any equation variable bound to these flow properties determine the value of a variable bound to an *out*-flow property, then eliminate the *inout*-flow property and give a new constant value to its variable by binding to a *PhSConstant*-stereotyped property that has a constant default value or instance value (see [14] for value assignment examples). If any equation variable bound to these flow properties is determined by a variable in the same equation that is bound to an *in*-flow property, then eliminate the *inout*-flow property and give its variable a new binding to a new property with a *PhSVariable* (see [14] on applying variable- and constant-value stereotypes to properties in SysPhS).

The remaining sections present the debugging techniques. Static techniques find the cause of failures to compile and simulate translated models. This type of failure prevents generating a simulation run-time from the translated model. Once compilation succeeds, dynamic debugging techniques identify causes of failure to produce intended simulation behavior. The underlying theme for static debugging is tracing symbolic transformations in the model to find errors. Transformation tracing is also useful for dynamic debugging to better understand the model and sources of potential simulation-related errors.

B. *Static Debugging for Failure to Execute Simulation*

The failure of an equation-based model (translated from a system model) to compile and simulate in a simulation tool indicates a static error. These errors can be identified with static debugging techniques applied to the system model, which trace chains of symbolic transformations in the model. These transformations appear as mathematical relationships in constraint equations or implied by connectors. Specifically, tracing refers to tracking transformations of known and unknown variables through a model. Known variables are properties whose values are assigned a constant value or determined through mathematical relationships. Tracing is complemented by bookkeeping, which records the known or unknown status of these variables when operations apply to them in the model.

Static debugging can be performed on complete system models, but is described here on simplified, complementary models of a system's physical interactions and signal flows. For models with physical interactions, the first task is to identify the part, port, or connector property in IBDs where physical interaction will first occur or initiate other physical interactions in the system. Multiple parts and ports where physical interactions simultaneously occur can initiate further interactions, but any one can be arbitrarily picked to begin tracing. Tracing and bookkeeping of mathematical transformations start with properties associated to this selected part or port. Deciding which system component commences the physical interactions is easy in many cases. For example, the initiators of flow of electric charge in an electric circuit are the voltage sources or current sources. In the cruise control system represented in IBDs in Figures 1 and 2, the throttle in the engine physically initiates the car's interaction with the road and air (this happens on command from the driver, but the command is signal flow, not physical interaction).

When the initiator of physical interaction is not obvious, it can help to inspect the parametric diagrams of parts or ports in IBDs. Parametric diagrams contain bindings between

Dadfarnia, Mehdi; Barbau, Raphael.
"Platform-Independent Debugging of Physical Interaction and Signal Flow Models."
Paper presented at The 13th Annual IEEE International Systems Conference, Orlando, FL, United States. April 8, 2019 - April 11, 2019.

properties of the parts (or ports) to variables in the part's constraint equations. Look for parametric diagrams of parts that have a higher number of *PhSConstant*-stereotyped properties (with values given explicitly in the model) than *PhSVariable*-stereotyped properties (with values determined by mathematical relationships in the model), except for *PhSVariables* that give simulation time. The equation variables (constraint parameters) bound to *PhSConstant* or time properties are used in the part's equations (constraints) to determine values of other variables, which are bound to other properties used in the part's equations. To find an initiator, search for a part or port where most of its properties or properties of its ports are bound to constants or time values in its parametric diagram. The only properties without constant or time values should be flow properties, which can only have their values determined through connectors. Parts or ports initiating physical interactions have the fewest of these flow properties.

Tracing bindings and constraints in parametric diagrams helps understand and keep track of (bookkeep) which variables in the equations are known and unknown. Constraint equations show mathematical transformations between known variables, bound to properties with known values, and unknown variables, bound to properties with unknown values. Before simulation, the only known variables are the ones bound to *PhSConstant* properties, the variables bound to properties given (initial) values at the start of simulation, and properties that give simulation time values. These should lead to values assigned to all variables in the parametrics diagram of physical interaction-initiating parts. The status of these variables will change as tracing shows their values being assigned through constraints or connectors, which is recorded by bookkeeping.

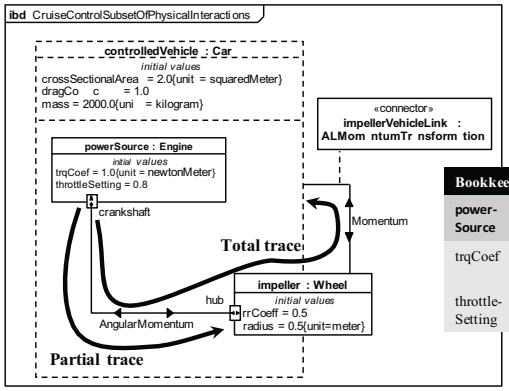
Physical interaction flow properties on the current part in the debugging process link to flow properties on parts or ports at the other end of the linking connectors. Trace along these connectors to find out whether values are assigned to these flow properties leads to parametric diagrams of other parts, ports, and connector properties linked to the current part. Repeat the same methods of tracing and bookkeeping in these other parametric diagrams to determine whether values are assigned to unknown variables and to find flow properties that lead to new connectors and parametric diagrams. The trace must go through all connectors and parametric diagrams of the system's parts, ports, and connector properties. Figure 4 shows an example of tracing and bookkeeping value assignments between a physical interaction-initiating part and another part. Bookkeeping of the total trace would complete the tracking of value assignments.

A system model will compile and simulate when translated if it a) uses all the constraint equations and connectors in the model for mathematical transformations between known and unknown variables and b) has all its property values determined by simulation of mathematical transformations. If tracing and bookkeeping identifies a constraint equation or connector that is not used, the system is overconstrained. In this scenario, the modeler must choose whether unused equations or connectors should be removed or a new property should be included and related to them. If an unknown property is not defined by any mathematical constraint or connector, then the system is underconstrained. In this scenario, the modeler must choose between using this property in a new equation or eliminating the property. Tracing and bookkeeping of equations also helps spot constraint equations that involve a division by zero and functions called outside their domains. Once corrections to the model are made, they are replicated in the original system model.

If there is a complementary model of signal flows, repeat the process of tracing and bookkeeping in a similar fashion, but start tracing from all parts that do not have *in*-flow properties or do not own ports that have *in*-flow properties. The *in*-flow property on these parts indicate that they receive unidirectional signals from another part in the model, so they cannot be the initiator of signal flows. Corrections in this model should likewise be reproduced in the original, complete model of the system. Translate the corrected SysML model and test on simulation platforms to determine if more debugging is needed.

C. Dynamic Debugging for Unexpected Simulation Results

Failure of an equation-based simulation model (translated from a system model) to produce expected results when executed indicates a dynamic error. The simulation model is able to compile and simulate, but produces variable values that deviate from modeler expectations. These errors can be identified with dynamic debugging techniques applied to



Bookkeeping of variables through parts and ports in the partial trace

power-Source	Value known?	crankshaft	Value known?	hub	Value known?
trqCoef	☒	torque	☒	torque	☒
throttle-Setting	☒	angular velocity	☒	angular velocity	☐

Fig.4. Shows initiating physical interaction component, direction of traces, bookkeeping of variables, and value assignment that occur through the partial trace (for brevity)

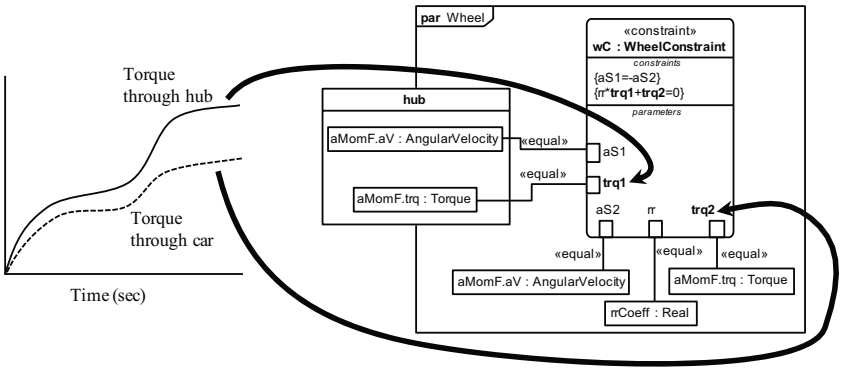


Fig. 5. Relationship between simulation variables and flow properties in the system IBD

the system model. These techniques examine executed simulations to understand exactly when signals and conserved substances flow through the system and what their characteristics are. They focus on simulation results for variables involved in the static traces of flow properties linked by connectors in the previous section. This showed how variables characterizing flow of physical substances and signals during simulation are related via transformations in the system model (mathematical operations via constraint equations and connectors). Though dynamic debugging can be performed without prior static debugging, fixing static errors first ensures the simulation model will compile and execute, and static tracing improves understanding of how variables change during simulation.

Dynamic debugging can be performed on complete system models, but is described here on simplified, complementary models of a system's physical interactions and signal flows. Behavior of conserved substances in physical interactions is characterized by their flow rate and potential to flow. Flow rate and potential to flow appear in simulation as variables translated from properties at the ends of connectors in the system model. This enables modelers to track simulation variables that correspond to properties in SysML system models. The SysPhS translator uses the names of association ends and constraint parameters in the resulting simulation models to facilitate this, but tracking simulation variables might require some familiarity with the equation-based simulation language. Lastly, like static debugging, dynamic debugging starts by tracing simulation variable transformations at points in the model that initiate physical interactions in the rest of the model. These points must be identified before debugging.

Physical interaction variables simulate flow of conserved substances only at their corresponding connector endpoint (part or port) in the system model. A more complete picture of symbolic transformations of these variables is seen by observing their values over simulated time and comparing them to other physical interaction simulation variables in the model. Graphical displays in simulation tools show these values, enabling comparison of simulated values to their intended mathematical relationships. The relationships are defined, correctly or not, through transformations (mathematical relationships between variables derived from connectors and parametric diagrams in the system model) of corresponding flow properties in the system model. To visualize these transformations, observe variables when their corresponding flow properties have not undergone more than one set of transformations (operations that occur on flow properties in the constraints of one parametric diagram or in the mathematical relationship implied by one connector). Compare simulation values of these variables with those of other physical interaction variables related to the same part or port in the system model, as well as simulation variables related to the other end of the variables' associated connectors.

Analysis of simulation variable results is performed in simulation runs that are sufficiently long for their values to reach a steady-state or a recognizable pattern of changes. Check that changes follow the mathematical transformations specified in corresponding constraint equations and connector links in the system model, which can be modified to produce better results. Figure 5 shows the relationship between simulated variable values over time and flow properties in the parametrics diagram and IBD.

Further simplification of system models can determine whether simulation results are valid, especially when physical interactions are highly complex. One way is to temporarily remove parts, ports, and connectors until modelers have high confidence in what they expect from variable behavior. Once this core model produces correct simulations, the removed parts and ports can be incrementally restored, simulated, and checked [16].

Validity of simulation results might also be determined by reaching consensus among modelers, users, and stakeholders on whether the simulation model is producing the correct results [25]. These techniques are out of the scope of this paper, but they include qualitative and quantitative model exploration, and comparison of simulation results to system behavior or alternative validated simulation results [25].

If there exists a complementary model of signal flows, repeat the process of inspecting simulation variables in a similar fashion. However, start tracing with all parts that do not have *in*-flow properties or do not own ports that have *in*-flow properties, as chosen during static debugging. Replace remaining parts in a complementary model of signal flows that only have *out*-flow properties or only have ports with *out*-flow properties have their flow properties by *PhSConstant*-stereotyped properties with pre-specified values before debugging.

Errors that are found by debugging are corrected in the system model, then tested by translating to simulation models and executing them. Translating and testing system models to multiple simulation platforms is more robust, because fixes

sometimes work for one simulation platform and not others. For example, a function call in a parametric diagram is domain-specific, and this might need to be replaced with a more universal function call. It is also possible that some modeling capabilities in SysML, such as state machines or different ways of defining initial values, cannot be replicated on some simulation platforms (see [14] for more specific examples about translation differences between simulation platforms).

IV. Conclusions & Future work

This paper presents an overview of debugging procedures for physical interaction and signal flow models translated from system models to equation-based simulation languages. The procedures identify errors causing compilation and simulation of these models to fail, or to produce incorrect simulation results. The integration of system models with equation-based models facilitates interoperability between developers of these types of models. This is done in SysML extended by SysPhS [6, 12, 14, 15]. The debugging procedures for platform-independent system models help identify problems without debugging and correcting the domain-specific simulation model and then transferring those changes back into the system model. The procedures are categorized as static and dynamic. Static debugging traces symbolic transformations in the system model, and dynamic debugging uses results of simulations to check changes in variable values during simulation. These debugging procedures are performed on system models and intend to complement existing debugging techniques on simulation platforms.

The authors plan to improve the debugging processes with user-friendly interfaces to visualize aspects of model translation, particularly the mapping between components in system models (e.g. equations, parts, properties) and structures in simulation models resulting from translation. The interface can provide information to both system and simulation modelers about the names, locations, and number of times translated structures and variables appear in the simulation model. This facilitates communication between systems engineers and simulation tool experts, without being concerned about details in one another's models.

Acknowledgments

The authors thank Conrad Bock for helpful discussions on the contents of this paper. Commercial equipment and materials might be identified to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the U.S. National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] A. Pop, M. Sjölund, A. Asghar, P. Fritzson, and F. Casella, "Integrated Debugging of Modelica Models," *Modeling, Identification and Control,* vol. 35, no. 2, pp. 93-107, 2014.

[2] S. Friedenthal, A. Moore, and R. Steiner, *A practical guide to SysML: the systems modeling language*, Morgan Kaufmann Publishers, 2014.

[3] A. L. Ramos, J. Vasconcelos Ferreira, and J. Barceló, "Model-based systems engineering: An emerging approach for modern systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 1, pp. 101-111, 2012.

[4] P. Fritzson, *Introduction to modeling and simulation of technical and physical systems with Modelica*, John Wiley & Sons, 2011.

[5] S. Van Mierlo, "A multi-paradigm modelling approach for engineering model debugging environments," Ph.D. dissertation, Department of Mathematics and Computer Science, University of Antwerp, 2018.

[6] C. Bock, R. Barbau, I. Matei, and M. Dadfarnia, "An Extension of the Systems Modeling Language for Physical Interaction and Signal Flow Simulation," *Systems Engineering* vol. 20, no. 5, pp. 395-431, 2017.

[7] Controllab Products B.V., *Getting Started with 20-Sim 4.6*, 2017. Accessed on: Oct. 16, 2018. [Online]. Available: http://www.20sim.com/downloads/files/20simGettingStarted46.pdf

[8] F.E. Cellier, *Continuous System Modeling*, Springer Science & Business Media, 1991.

[9] The MathWorks, Inc., *Simscape Language Guide*, *Version 4.5*, September 2018. Accessed on October 18, 2018. [Online]. Available: https://www.mathworks.com/help/pdf_doc/physmod/simscape/simscape_lang.pdf

[10] Open Source Modelica Consortium, *OpenModelica User's Guide*, February 18, 2016. Accessed on October 18, 2018. [Online]. Available: https://www.openmodelica.org/doc/OpenModelicaUsersGuide/OpenModelicaUsersGuide-v1.9.4-dev.beta2.pdf

[11] *IEEE Standard for Application and Management of the Systems Engineering Process,* IEEE 1220-2005, 2005.

[12] M. Dadfarnia, C. Bock, and R. Barbau, "An Improved Method of Physical Interaction and Signal Flow Modeling for Systems Engineering," in *14th Annual Conference on Systems Engineering Research (CSER),* Huntsville USA, March 2016.

[13] Object Management Group, *OMG Systems Modeling Language Specification, version 1.5*, May 2017. Accessed on October 19, 2018. [Online]. Available: http://www.omg.org/spec/SysML/1.5

[14] Object Management Group, *SysML Extension for Physical Interaction and Signal Flow Simulation Specification, Version 1.0,* July 2018. Accessed on October 2, 2018. [Online]. Available: https://www.omg.org/spec/SysPhS/1.0/

[15] R. Barbau, *SysPhS 1.0 OMG Release*, May 4, 2018. GitHub repository, accessed on November 2018. [Online]. Available: https://github.com/usnistgov/saismo/releases/tag/sysphs

[16] D. Krahl, "Debugging simulation models," in *37th Conference on Winter Simulation*, Olrando USA, December 2005, pp. 62-68.

[17] Modelica Association, *Modelica-A Unified Object-Oriented Language for Systems Modeling, Language Specification, Version 3.3, Revision 1*, July 2014. Accessed on October 18, 2018. [Online]. Availabe: https://www.modelica.org/documents/ ModelicaSpec33Revision1.pdf

[18] P. Bunus, and P. Fritzson, "A debugging scheme for declarative equation based modeling languages," in *International Symposium on Practical Aspects of Declarative Languages*, Portland USA, January 2002, pp. 280-298.

[19] P. Bunus, and P. Fritzson, "Semi-automatic fault localization and behavior verification for physical system simulation models," in *18th IEEE International Conference on Automated Software Engineering,* Montreal Canada, October 2003, pp. 253-258.

[20] A. Pop, and P. Fritzson, "A portable debugger for algorithmic modelica code," in *4th International Modelica Conference,* Hamburg Germany, March 2005, pp. 435-443.

[21] A. Asghar, A. Pop, M. Sjölund, and P. Fritzson, "Efficient Debugging of Large Algorithmic Modelica Applications," *IFAC Proceedings Volumes* vol. 45, no. 2, pp. 1087-1090, 2012.

[22] M. Sjölund, F. Casella, A. Pop, A. Asghar, P. Fritzson, W. Braun, L. Ochel, and B. Bachmann, "Integrated Debugging of Equation-Based Models," in *10th International Modelica Conference,* Lund Sweden, March 2014, pp. 195-204.

[23] M. Sjölund, and P. Fritzson, "Debugging symbolic transformations in equation systems," in *4th International Workshop on Equation-Based Object-Oriented Modeling Languages and Tools,* Zurich, Switzerland, September 2011, pp. 67-74.

[24] A. Canedo, and L. Shen, "Functional Debugging of Equation-Based Languages," in *5th International Workshop on Equation-Based Object-Oriented Modeling Languages and Tools,* Nottinham UK, April 2013, pp. 55-64.

[25] R. G. Sargent, D. M. Goldsman, and T. Yaacoub, "A tutorial on the operational validation of simulation models," in *Winter Simulation Conference (WSC),* Washington D.C. USA, December 2016, pp. 163-177.

Dadfarnia, Mehdi; Barbau, Raphael.
"Platform-Independent Debugging of Physical Interaction and Signal Flow Models."
Paper presented at The 13th Annual IEEE International Systems Conference, Orlando, FL, United States. April 8, 2019 - April 11, 2019.

Measuring Temperature Distribution in Steel-Concrete Composite Slabs Subjected to Fire Using Brillouin Scattering Based Distributed Fiber Optic Sensors

Yi Bao[1], Matthew S. Hoehler[2], Christopher M. Smith[3], Matthew Bundy[2], Genda Chen[4]
[1] *Stevens Institute of Technology – USA, Email:* yi.bao@stevens.edu
[2] *National Institute of Standards and Technology - USA,* [3] *Berkshire Hathaway Specialty Insurance - USA,* [4] *Missouri University of Science and Technology - USA*

Abstract

This study investigates temperature distributions in steel-concrete composite slabs subjected to fire using distributed fiber optic sensors. Several 1.2 m × 0.9 m composite slabs instrumented with telecommunication-grade single-mode fused silica fibers were fabricated and subjected to fire for over 3 hours. Temperatures were measured at centimeter-scale spatial resolution by means of pulse pre-pumped Brillouin optical time domain analysis. The distributed fiber optic sensors operated at material temperatures higher than 900 °C with adequate sensitivity and accuracy to allow structural performance assessment, demonstrating their effective use in structural fire applications. The measured temperature distributions indicate a spatially-varying, fire-induced thermal response in steel-concrete composite slab, which can only be adequately captured using approaches that provide high data point density.

1. Introduction

The mechanical properties of construction materials and the load-carrying capacity and stability of structural members (beams, columns, slabs, and joints) are reduced at elevated temperatures (Li and Zhang 2012; Li et al. 2017a-c). To understand structural behavior during and after a fire when measured material temperature histories are unavailable, coupled thermo-mechanical analyses are often conducted (Jeffers and Sotelino 2012; Bao et al. 2016). A great number of analytical and numerical approaches have been developed to predict gas-phase temperature distributions and evolution histories resulting from fire. Zone models (Li and Zhang 2012), computational fluid dynamics models (McGrattan et al. 2010) or stochastic models (Bertola and Cafaro 2009) are often used. To date, it remains a challenge to accurately predict temperature distributions in structural members through the heat transfer analysis based on gas temperatures and radiative heat flux, in particular for structures with complex geometry, such as composite floors. The error in the predicted structural member temperature distribution and evolution over time can result in inaccurate conclusions about the mechanical response of the structure. Additionally, the uncertainty in structural element temperature distribution cannot easily be quantified.

Bao, Yi; Hoehler, Matthew; Smith, Christopher; Bundy, Matthew; Chen, Genda.
"Measuring Temperature Distribution in Steel-Concrete Composite Slabs Subjected to Fire Using Brillouin Scattering Based Distributed Fiber Optic Sensors."
Paper presented at 9th International Conference on Structural Health Monitoring of Intelligent Infrastructure, St. Louis, MO, United States.
August 4, 2019 - August 7, 2019.

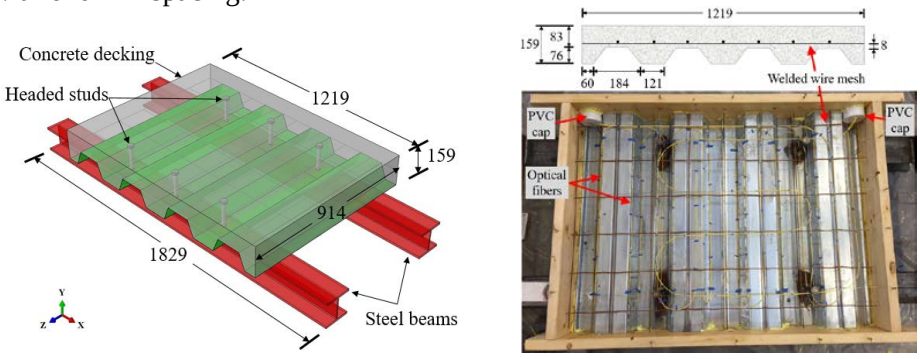
If temperatures in structural members during a fire can be measured with fine spatial resolution, understanding the structural response due to thermal loading becomes more tractable. Traditionally in structural fire research, temperature is measured using thermocouples deployed at a limited number of discrete points. Recently, fully-distributed fiber optic sensors using pulse pre-pumped Brillouin optical time domain analysis (PPP-BOTDA) have successfully been used to measure temperature and strain in various structural applications (Bao and Chen 2016 a, b). Each instance of PPP-BOTDA can provide hundreds of temperatures measurements along the length of an optical fiber.

In this study, telecommunication-grade single-mode optical fibers are embedded in profiled concrete slabs to measure temperature distributions under fire condition. The obtained temperature data can be used to understand the structural behavior of the composite slabs.

2. Experimental Program

2.1 Specimen

The composite specimens were fabricated to develop installation procedures for optical fibers in steel-concrete composites and investigate the response of the optical fibers during a fire. For brevity, only one representative slab is presented in this paper. The specimen was a reinforced concrete slab supported on two W5×19 steel beams as depicted in Fig. 1(a). Composite action between the concrete slab and steel beams was achieved using headed steel studs. The concrete slab was 1219 mm long and 914 mm wide. It was cast on 0.9 mm thick, trapezoidal metal decking (Vulcraft 3VLI20[1]). The depth of the concrete slab varied from 83 mm to 159 mm as illustrated in Fig. 1(b). The concrete slab was reinforced with welded wire mesh (6×6 W1.4/W1.4), which has a specified mesh spacing of 150 mm and a wire diameter of 3.4 mm. The headed steel studs (Nelson S3L) had a specified shaft diameter of 19 mm and an effective embedment length of 117 mm. The two W5×19 steel beams were 1,829 mm long and placed in parallel with 610 mm spacing.



(a) (b)

Fig. 1. Steel-concrete composite slab specimens: (a) isometric rendering, and (b) cross-sectional and top views (all units in mm).

[1] Certain commercial products are identified in this paper to specify the materials used and the procedures employed. In no case does such identification imply endorsement or recommendation by the National Institute of Standards and Technology, nor does it indicate that the products are necessarily the best available for the purpose.

2.2 Material Properties

The design concrete mix used in this study was 0.193: 0.105: 0.315: 1.0: 1.136 by weight for water: fly ash: Type I Portland cement: river sand: expanded slate lightweight aggregate (LWA). This mix corresponded to a water-to-binder ratio of 0.46, a binder-to-sand ratio of 0.42, and a sand-to-LWA ratio of 0.88. The river sand had a diameter up to 4.75 mm. The LWA used had low water-retention characteristics and high desorption (Meng et al. 2017) to expedite the curing of the specimens. A polycarboxylate-based high-range water reducer was used to improve flowability of the concrete. Polypropylene microfibers (FRC MONO-150) with a nominal diameter of 38 μm and lengths between 13 mm and 19 mm were added to the mix at a dosage of 2.37 kg/m^3 of concrete to reduce thermal spalling.

During composite specimen casting, ten standard cylinders measuring about 102 mm in diameter and 203 mm in height were prepared for concrete material testing. Five cylinders were tested and analyzed to determine the average ± standard deviation of each concrete property. Specifically, the measured concrete density was (2,070 ± 80) kg/m^3 at 28 days. The compressive strength of concrete was (38 ± 3) MPa at 28 days and (41 ± 3) MPa at 56 days, which were determined in accordance with the American Society for Testing and Materials standard ASTM C39/C39M. A relative humidity sensor (Vaisala HM40S RH Probe) was inserted into each concrete slab from its top to measure the internal relative humidity of concrete at a depth of 90 mm.

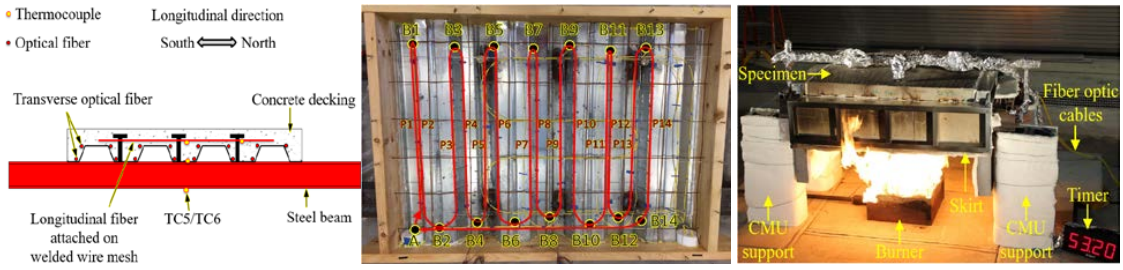
2.3 Instrumentation and Test Setup

The specimen was instrumented with three distributed fiber optic sensors as shown in Fig. 2. The result from one sensor is reported in this paper. The sensor was laid on top of the metal decking and ran parallel to the flutes of the metal decking. The curved portions of the fiber optic sensor were labeled as B1 to B14. The distributed sensor entered the concrete from a polyvinyl chloride (PVC) cap at Point A. The cap measured 100 mm in diameter and was used to protect the fibers from damage during concrete casting. The optical fiber had a buffer (diameter: 900 μm), an outer coating (outer diameter: 242 μm), an inner coating (outer diameter: 190 μm), a glass cladding (outer diameter: 125 μm), and a glass core (diameter: 8.2 μm) (Huang et al. 2013). The optical fiber was free to slide in the sheath so that it was approximately free of axial strain. Thus, the distributed sensor can be used to measure temperature changes (Bao et al. 2017).

Each composite specimen was also instrumented with six glass-sheathed, bare bead, K-type thermocouples (24-gauge wire). The thermocouples were designated TC1 to TC6: TC1 on the top surface of the metal decking in the center of the specimen, TC2 on the welded wire mesh in the center of the specimen, TC3 on the welded wire mesh 305 mm away from the mid-span, TC4 on a headed stud 300 mm away from the mid-span, and TC5 and TC6 peened into a small drill hole (1.5 mm) in the center bottom flanges of the steel beams. An Inconel-shielded thermocouple located 25 mm below the metal decking at the center of the compartment was used to measure the gas temperature below the concrete deck. The thermocouples have a manufacturer-specified temperature standard limit of error of 2.2 °C or 0.75 % (whichever value is greater) over a range of 0 °C to 1250 °C. The total expanded uncertainties for the material temperature and gas temperature measurements are ±6.2 % and ±14.7 % of the reading, respectively. The total

expanded uncertainty for the burner heat release rate is less than ±2.4 % (Bundy et al 2007). Data from the fuel delivery system and thermocouples were recorded at a rate of 1 Hz.

Fire tests were conducted in the National Fire Research Laboratory at the National Institute of Standards and Technology (NIST). The test setup was not intended to represent a particular structure, but rather to investigate the performance of fiber optic sensors in a typical steel-concrete composite structure. Fig. 3 depicts the test setup located under a 6 m × 6 m (plan) exhaust hood (not shown in photo). The flame source was a natural gas diffusion burner measuring 530 mm × 530 mm × 200 mm (length × width × height). Natural gas entered the burner near the bottom, filled the burner cavity and percolated through a gravel layer to distribute the gas. The burners were manually regulated using a needle valve on the gas supply line. A skirt constructed of cold-formed steel framing and cement board lined with thermal ceramic fiber blanket partially enclosed the space above the burner to trap hot gases beneath the specimen. The skirt was open at the bottom, creating the compartment fire dynamics depicted in Fig. 4(a). The heated 'compartment' created by the skirt was approximately 1220 mm × 920 mm × 30 mm (length × width × height). Each beam was simply-supported at a clear-span of 1530 mm on four supports made of stacked concrete masonry units (CMU). The supports were wrapped with 25-mm thick thermal ceramic blankets for additional thermal protection. Each specimen was subjected to four fire sizes. No mechanical loading beyond the specimen self-weight was applied. The magnitude of the fire was controlled through the burner's calculated heat release rate (HRR). The HRR was held approximately constant at four target levels: 50 kW, 100 kW, 150 kW, and 200 kW. Each of the first three levels was maintained for 2 min, and the last level (200 kW) was maintained for 210 min, before the fire was extinguished. In total, the specimen was heated for 216 min (3 h and 36 min). The compartment upper layer gas temperature corresponding to the HRR of 200 kW was sustained at about 900 ºC.



(a) (b) (c)

Fig. 2. Instrumentation layout: (a) overview, (b) top view of the distributed sensor, (c) test set-up.

3. Results and Discussion

Temperature distributions along the distributed sensor are plotted in Fig. 3. The time is relative to burner ignition. The horizontal axis represents the distance along the distributed sensor. The vertical axis represents the measured temperature, which was obtained from the Brillouin frequency shift measured by the distributed sensor and the frequency-temperature calibration curve (Bao et al. 2015). Temperature increases with time as expected during the heating phase of the experiment. The peaks in the temperature distributions in Fig. 3 are marked as 'Pn', where 'n' indicates the location of the peak corresponding to the positions shown Fig. 2(b). The first 14

peaks are marked as P1 to P14, which are along the centerline of the specimen. The fact that the peaks occurred along the centerline suggests that: (1) the gas temperature and radiative heat flux was lower near the edges of the test setup; (2) there was more heat loss from the specimen to the surrounding environment at its edges; and (3) the steel beams near the edges provided heat sink for the concrete slab. The temperature variation transversely across the specimen is significant; over 600 °C variation for fiber section B7 to B8. This spatial variation would commonly be neglected in thermo-mechanical analysis where it is typically assumed that gas temperature and heat flux below the slab is uniform. P8 and P9 exhibited the highest temperatures directly above the burner. Overall, the temperatures at P1, P4, P5, P8, P9, P12, and P13 are higher (on the order of 200 °C to 400 °C) than the temperatures at P2, P3, P6, P7, P10, P11, and P14, suggesting that the lower flanges of the concrete slab were subjected to more intense thermal conditions than the higher flanges because they were closer to the burner. The peaks after P14 are marked by the locations of the curved portions of the distributed sensor. For instance, the peak B10, which is a peak after P14, corresponds to the distributed sensor near B10.

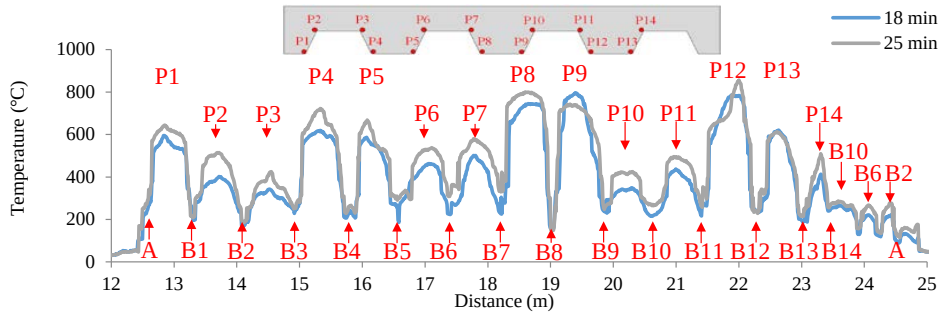


Fig. 3. Temperature distributions measured from the distributed fiber optic sensor.

4. Conclusions

In this study, pulse pre-pumped Brillouin optical time domain analysis (PPP-BOTDA) was used to measure temperatures in distributed fiber optic sensors installed in a steel-concrete composite slab specimen exposed to fire. The limited set of data demonstrated that the investigated commercially-available, polymer-sheathed optical fiber survived the concrete casting process. Material temperatures higher than 900 °C were measured at the interface between the concrete and the metal deck with adequate sensitivity and accuracy for typical structural engineering applications.

The measured temperatures from a distributed fiber optic sensor indicate highly non-uniform temperature distribution in the composite slab, which is often neglected in engineering design and analysis. Deploying the distributed fiber optic sensors in large-scale structural fire tests has potential to improve our understanding on the performance of infrastructure in fires and thus fire safety.

Acknowledgements

This work was funded by the National Institute of Standards and Technology (NIST) [grant No. 70NANB13H183]. The authors thank Alana Guzetta of US Concrete and Dale Bentz of NIST for their assistance in the concrete design.

References

Bao, Y., Chen, G. (2015). "Fully-distributed fiber optic sensor for strain measurement at high temperature," *Proc. 10th Int. Workshop Struct. Health. Monit.*, Stanford, CA.

Bao, Y., Chen, G. (2016a). "Temperature-dependent strain and temperature sensitivities of fused silica single mode fiber sensors with pulse pre-pump Brillouin optical time domain analysis," *Mes. Sci. Tech.*, 27(6), 65101–65111.

Bao, Y., Chen, G. (2016b). "High temperature measurement with Brillouin optical time domain analysis," *Opt. Lett.*, 41(14), 3177–3180.

Bao, Y., Chen, Y., Hoehler, S.M., Smith, M.C., Bundy, M., Chen, G. (2016). "Experimental analysis of steel beams subjected to fire enhanced by Brillouin scattering-based fiber optic sensor data," *J. Struct. Eng.*, 143(1), 04016143.

Bao, Y., Hoehler, M.S., Smith, C.M., Bundy, M., Chen, G. (2017). "Temperature measurement and damage detection in concrete beams exposed to fire using PPP-BOTDA based fiber optic sensors," *Smart Mater. Struct.*, 26(10), 105034.

Bertola, V., Cafaro, E. (2009). "Deterministic-stochastic approach to compartment fire modeling," *Proc. R. Soc. London, Ser. A*, 465, 1029–1041.

Bundy, M., Hamins, A., Johnsson, E.L., Kim, S.C., Ko, G.H., Lenhert, D.B. (2007). "Measurements of heat and combustion products in reduced-scale ventilation-limited compartment fires," *NIST Technical Note*, 1483.

Huang, Y., Fang, X., Zhou, Z., Chen, G., Xiao, H. (2013). "Large-strain optical fiber sensing and real-time FEM updating of steel structures under the high temperature effect," *Smart Mater. Struct.*, 22(1), doi:10.1088/0964-1726/22/1/015016.

Jeffers, A.E., Sotelino, E.D. (2012). "An efficient fiber element approach for the thermo-structural simulation of non-uniformly heated frames," Fire Safety J., 51, 18–26.

Li, G., Zhang, C. (2012). "Simple approach for calculating maximum temperature of insulated steel members in natural-fires," *J. Constr. Steel Res.*, 71, 104–110.

Li, X., Wang, J., Bao, Y., Chen, G. (2017a) "Cyclic behavior of damaged reinforced concrete columns repaired with environment-friendly fiber-reinforced cementitious composites," *Eng. Struct.*, 136, 26–35.

Li, X., Bao, Y., Xue, N., Chen, G. (2017b). "Bond strength of steel bars embedded in high-performance fiber-reinforced cementitious composite before and after exposure to elevated temperatures," *Fire Safety J.*, 92, 98–106.

Li, X., Bao, Y., Wu, L., Yan, Q., Ma, H., Chen, G., Zhang, H. (2017c). "Thermal and mechanical properties of high-performance fiber-reinforced cementitious composites after exposure to high temperatures," *Constr. Build. Mater.*, 157, 829–838.

McAllister, T., Luecke, W., Iadicola, M., Bundy, M. (2012). "Measurement of temperature, displacement, and strain in structural components subject to fire effects: concepts and candidate approaches," *NIST Technical Note*, 1768, 73.

Meng, W., Khayat, K.H. (2017). "Effects of saturated lightweight sand content on key characteristics of ultra-high-performance concrete," *Cem. Concr. Res.*, (101), 46–54.

Proceedings
9th International Conference on Structural Health Monitoring of Intelligent Infrastructure
August 4-7, 2019 – St. Louis, Missouri (USA)

Bao, Yi; Hoehler, Matthew; Smith, Christopher; Bundy, Matthew; Chen, Genda.
"Measuring Temperature Distribution in Steel-Concrete Composite Slabs Subjected to Fire Using Brillouin Scattering Based Distributed Fiber Optic Sensors."
Paper presented at 9th International Conference on Structural Health Monitoring of Intelligent Infrastructure, St. Louis, MO, United States.
August 4, 2019 - August 7, 2019.

Building Design Considerations to Support Immediate Occupancy Performance Objectives

Siamak Sattar[1], Christopher L. Segura Jr.[1], Katherine J. Johnson[2], Therese P. McAllister[3], and Steven L. McCabe[4]

[1] Research Structural Engineer, National Institute of Standards and Technology (NIST), Gaithersburg, MD USA.
[2] Earthquake Risk Mitigation Policy Analyst, NIST, Gaithersburg, MD USA.
[3] Community Resilience Group Leader, NIST, Gaithersburg, MD USA.
[4] National Earthquake Hazards Reduction Program Director, NIST, Gaithersburg, MD USA.

ABSTRACT
The intent of current building codes for typical commercial and residential buildings is to safeguard against loss of life to building occupants by minimizing the probability of structural collapse during natural hazard events. However, preserving building functionality after a natural hazard is not the primary consideration in current codes. Widespread building damage, and degradation or loss of building functions, can have severe social and economic impacts on a community. To reduce the likelihood and severity of potential property damage and to enable more rapid recovery for communities impacted by natural hazards, the U.S. Senate tasked the National Institute of Standards and Technology (NIST) with identifying research needs and implementation activities to develop a multi-hazard immediate occupancy (IO) performance objective for commercial and residential buildings. This new IO performance objective would provide the technical criteria needed to design a building to retain its function following a design level hazard event. With input from subject matter experts and stakeholders participating in a national workshop, NIST developed a report to fulfill the Congressional mandate. This paper highlights key research needs and implementation activities that pertain to designing a building to meet IO performance objectives. The research needs and implementation activities discussed in this paper articulate the steps necessary to develop the engineering design criteria for the new IO performance objective. For successful development of IO objectives, these topics need to be addressed with cooperative efforts among researchers, engineers, standards and code officials, and community stakeholders.

Keywords: immediate occupancy, performance objective, building design, performance objective, adoption consideration.

INTRODUCTION
Current building codes mainly focus on preserving lives of building occupants, and generally do not address continued functionality after a hazard event. However, societal needs are quickly outpacing the performance goals for which current building codes have been developed. Communities, owners, and residents can benefit from buildings that are more resilient to natural hazard events to avoid lengthy and costly repairs or rebuilding, as well as minimizing the need for long-term evacuation of building occupants. To address this issue, the U.S. Senate directed the National Institute of Standards and Technology (NIST) to identify engineering principles, research, and implementation activities needed for a new "safety building performance objective for commercial and residential properties" [1].
In response to the congressional mandate, NIST developed a report identifying the research needs and implementation activities required to develop immediate occupancy (IO) performance objectives [2]. This report was developed through a collaborative process with a steering committee of subject matter experts and a national expert stakeholder workshop hosted by NIST. In the NIST report, IO performance is considered as the building's condition after a hazard event where damage to the building's structural system is controlled, limited, and repairable while the building remains safe to occupy. The building's ability to function at full or minimally reduced capacity is also affected by the functionality of the non-structural systems of the building (e.g., building envelope, equipment, interior utilities), as well as the infrastructure that connects the building to its surrounding community. Although other terminologies such as functional recovery may sound suitable to relay this concept, the term IO is used to highlight a potential range of functional levels, and for consistency with the congressional language. The role of lifelines in supporting the operation of functional buildings is acknowledged, but not addressed in detail in the NIST report. The NIST report covers improvements to building design, as well as community, economic and social, and adoption and acceptance considerations. This paper highlights research needs and implementation activities identified in the NIST report that pertain to designing individual buildings for IO performance.

Sattar, Siamak; Segura, Christopher; Johnson, Katherine; McAllister, Therese; McCabe, Steven.
"Building Design Considerations to Support Immediate Occupancy Performance Objectives."
Paper presented at 12th Canadian Conference on Earthquake Engineering, Quebec City, Canada. June 17, 2019 - June 20, 2019.

BUILDING DESIGN CONSIDERATIONS

There are significant challenges associated with designing new and retrofitting existing commercial or residential buildings to meet IO performance objectives. General challenges and research needs associated with designing a building for IO are presented within six subtopic areas: 1) Functionality Levels; 2) Damage Levels; 3) Design Practice; 4) Building Materials and Technologies; 5) Maintenance, Repair and Retrofit Methods; and 6) Monitoring and Assessment. Research and implementation activities that apply to each subtopic are presented in the proceeding section.

Functionality Levels

One of the primary differences between IO performance objectives and current building code design objectives is the consideration of the functionality level of a building following a hazard event. Because the post-hazard functionality of a building is not typically considered in the design process, building designers currently lack the tools and expertise necessary to assess a building's potential functionality for different loading scenarios, as well as the tools needed to determine the timeframe required for a building to return to its pre-hazard functionality level following a hazard event. Major challenges in incorporating functionality of the building to the design process include a general lack of understanding of the diverse factors that can impede building functionality, and the lack of adequate data to relate functionality and recovery time to building damage. Recommended research and implementation activities to achieve these goals are summarized below.

Functionality Classification: Building functionality classification levels that are appropriate to meet IO objectives need to be established. The functionality levels should describe the ability of the building to meet its intended purpose prior to a hazard, and to retain essential functions and return to pre-hazard functionality within a predetermined, acceptable timeframe following a hazard.

Recovery data: Detailed field reconnaissance data that describe the recovery process for buildings over time following hazard events are needed. The data can be used to inform the development of functionality levels and to develop and validate numerical models for predicting building functionality level. A standardized recovery data collection process should be developed to benefit ongoing efforts to develop and improve IO performance objectives for new and existing buildings.

Factors that influence functionality and recovery time: Research is needed to characterize the building systems that most influence building functionality, including those that are not commonly considered when planning and designing a building. This includes consideration of the impacts of damaged connections to utilities (e.g., water and electricity). In addition, the influence of maintenance and repair on building functionality needs to be better understood.

Analytical tools to predict building functionality: To enable the adoption of new immediate occupancy performance objectives that directly consider post-hazard functionality, building designers would need to be equipped with analytical tools to assess functionality for a variety of hazard scenarios. Building upon findings of the basic and applied research recommended within the Functionality Levels subtopic, analytical tools should be developed to quantify the functionality of a building over time, considering the impacts of hazard damage, degradation of building components, and maintenance, repair, and retrofit. The analytical tools should be developed in a reliability-based format using quantifiable data, from which a level of uncertainty can be expressed for achieving desired functionality levels.

Damage Levels

The functionality of a building depends on the amount of damage experienced by the building as a result of hazard events. In an engineering framework, building functionality needs to be measured as a function of damage to the building's structural components, nonstructural components, and any other components or equipment that can hinder functionality. To develop IO performance objectives, damage levels that are acceptable for IO objectives need to be identified, and relationships that link damage and functionality need to be developed. One of the key challenges of damage quantification is improving understanding of building response under different hazard types. Research to address this challenge includes experimental and field studies to characterize the performance of nonstructural components of a building. In addition, more accurate and simple numerical modeling techniques are needed to simulate the damage response of structural and nonstructural components within a building, including building contents, as well as their interaction. Recommended research and implementation activities to achieve these goals are summarized below.

Acceptable damage levels: Research is needed to quantify acceptable levels of damage for each of the immediate occupancy functionality levels at the component level as well as at the system level.

Field reconnaissance and laboratory data: Detailed field reconnaissance and laboratory data are needed to characterize the damage types and quantify damage levels for individual components as well as the global performance of a building. The data

Sattar, Siamak; Segura, Christopher; Johnson, Katherine; McAllister, Therese; McCabe, Steven.
"Building Design Considerations to Support Immediate Occupancy Performance Objectives."
Paper presented at 12th Canadian Conference on Earthquake Engineering, Quebec City, Canada. June 17, 2019 - June 20, 2019.

can be used to support the development of appropriate functionality levels for immediate occupancy and to develop and validate damage prediction tools. Reconnaissance data should reflect the performance of a large number of buildings of various age and construction type, with various levels of damage, including buildings that sustained little or no damage.

Understanding damage levels for different hazards: Research is needed to improve the understanding of building response both at the component and systems levels under different hazard types and levels. In addition to studying individual building components, the interactions among and between structural and nonstructural components need to be studied to identify how these responses can affect one another.

Understanding degradation due to aging and environmental effects:
There is a substantial need to understand how buildings and their structural and non-structural components age and respond to environmental factors, such as exposure to ultraviolet radiation, humidity, and temperature. This effort includes experimental as well as field studies to characterize the aging and environmental effects for materials used in new and existing buildings. Research is needed to develop numerical models capable of simulating the impacts of environmental factors on the response and functionality of a building.

Damage prediction models: Research is needed to improve the accuracy of damage prediction models, and to develop numerical models for different materials and components that accurately simulate damage formation and propagation. This effort includes developing physics-based models as well as simplified models that can predict the damage response under different hazard types. The impact of cumulative damage from multiple hazard events needs to be investigated. Development of damage prediction models should include evaluating the impact of secondary sources of damage. Improved damage prediction models will inform the design, repair, and retrofit process, as well as the functionality level and recovery time of buildings.

Standardize data collection:
One of the challenges in analyzing data from different reconnaissance studies is the lack of standardization and interoperability among the datasets. Different teams may collect different types of data and use different collection protocols. Research is needed to identify the crucial data that need to be collected in reconnaissance studies. Research is also needed to minimize the human bias in the collected data from field studies. Protocols and guidelines on sampling and data collection in reconnaissance studies should to be developed to improve the consistency in the data collected by different teams and to ensure that data are recorded in a consistent and interoperable manner.

Design Practice

To design a building, engineers, architects, and building developers must evaluate the building and its systems by a set of technical design criteria that imply conformance with building code and standards requirements. Because buildings designed for IO performance will be required to meet damage and functionality criteria that are more restrictive than current prescriptive code requirements, the development of new guidelines and standards is necessary. These IO-specific design tools should: 1) characterize the hazard types, hazard levels, and hazard scenarios (e.g., multiple hazard events) to be considered for design; and 2) provide guidelines to utilize new design technologies that emphasize damage avoidance and repairability as needed to satisfy IO performance goals. To evaluate a building's conformance with these new design standards, the development of new analytical models that can reliably predict building damage and functionality levels is needed. To support the development of new IO design standards and analytical tools, research is needed to benchmark the performance anticipated for code-compliant buildings under various hazard scenarios. Recommended research and implementation activities to achieve these goals are summarized below.

***Hazard levels and considering multiple hazards in design*:** The hazard types and distinct hazard levels that should be considered in the IO design process need to be clearly articulated in the design criteria used by building designers. Improved hazard models and the development of hazard risk maps are essential for the development of IO design guidelines and standards. As hazard risk resources are developed, research should be conducted to express the appropriate hazard levels to evaluate the performance of buildings designed for IO performance. Individual buildings are often vulnerable to different hazard types and multiple occurrences of certain hazards over the building's lifecycle. Research is needed to determine the loading scenarios that are appropriate for buildings prone to various hazards, with consideration of their joint probability of occurrence, and to develop analytical tools and design guidelines for assessing the performance of damaged buildings.

***New design philosophies for immediate occupancy*:** A major challenge associated with immediate occupancy is the need to minimize building damage and the impacts of that damage. Research is needed to study the potential performance benefits of implementing new and existing low damage and rapidly reparable alternatives to common building designs. Numerical simulations and laboratory tests are needed to compare the reparability, functionality, and recovery timeframe for traditional

Sattar, Siamak; Segura, Christopher; Johnson, Katherine; McAllister, Therese; McCabe, Steven.
"Building Design Considerations to Support Immediate Occupancy Performance Objectives."
Paper presented at 12th Canadian Conference on Earthquake Engineering, Quebec City, Canada. June 17, 2019 - June 20, 2019.

building designs to those of low damage alternative for various hazard scenarios. For certain hazard events and certain buildings, damage may be unavoidable. Research is needed to identify methods to streamline the repair and recovery processes for buildings damaged in hazard events. Research investigating the ways in which repair and recovery strategies can be incorporated into the building design process is important. This research should investigate ways to optimize recovery by minimizing the quantity of damaged building components and designing rapid repair methods to be implemented if damage occurs.

***Benchmarking the performance of code-compliant buildings*:** Current building codes and standards are developed to ensure life safety of the occupants and to provide some degree of property protection; however, it is challenging to evaluate the reliability of those codes and standards to meet certain performance objectives (e.g., collapse prevention). These challenges are associated with, among other things: limited field data on the impacts of design level or extreme level hazards on code-compliant buildings; and capacity limitations of structural laboratories that make it difficult to conduct tests on large-scale building models. Nonetheless, research to benchmark the performance of code-compliant buildings should be prioritized, and a thorough evaluation of design standards for Risk Category III & IV buildings should be conducted to determine their suitability in meeting IO objectives.

***Data on the performance of existing buildings*:** Hazard response data from sensors in buildings and similar data from laboratory tests is needed to develop, calibrate, and validate numerical models that assess building performance. As much as possible, building information should summarize building component design parameters (e.g., geometry, materials) such that the hazard resistance of the building systems can be predicted using newly developed damage prediction and functionality prediction models.

***Analytical models to conduct building performance evaluations*:** The development of new analytical simulation tools, or modification of existing tools, is needed to incorporate newly developed damage prediction models and functionality prediction models into the building performance evaluation process. The newly developed tools should be capable of capturing the cumulative impacts of damage and aging on all structural and nonstructural building components, making it possible to evaluate the hazard performance and functionality of a building over its lifecycle. These tools should enable integrated modeling of all of a building's systems, accounting for interdependencies between the systems.

***Design requirements and performance evaluation criteria for immediate occupancy*:** New design guidelines and standards should be developed that dictate the technical criteria, hazard types, and hazard levels appropriate for designing a building for IO objectives. The guidelines should include hazard-specific performance criteria by which new and existing buildings are evaluated for their effectiveness in meeting IO safety and functionality requirements. These criteria should be informed by laboratory data, field data, and numerical studies, considering the appropriate level of risk for different damage levels, hazard types, and hazard levels. Because mechanical systems and envelope of a building play a crucial role in the ability of a building to conduct its intended functions, guidelines and standards for nonstructural building systems should be developed in a manner that is consistent with those developed for the structural system.

Building Materials and Technologies

New and improved building materials and technologies can offer improved options to prevent, detect, and mitigate building damage and expedite post-hazard recovery times. Research is needed to develop materials, technologies, and strategies that can decrease the likelihood of damage, lessen potential repair costs, and reduce potential recovery timeframes. Recommended research and implementation activities to achieve these goals are summarized below.

***Lower-cost materials*:** High materials costs can hinder decisions to repair damaged buildings and to retrofit buildings that are vulnerable to damage. The development of lower cost materials, both for the construction of new buildings and retrofit of existing buildings, must accompany the development of IO objectives.

***Performance of new materials*:** Research is needed to evaluate the performance of new materials when subjected to various hazards and environmental conditions. Research that investigates the economic impacts of using various building materials throughout a building's lifecycle, including post-hazard repairs, should be conducted to evaluate the potential long-term benefits of using new materials.

***Materials for modular, standardized construction and repair*:** Research is needed to develop new repair methods and modular building construction and modification techniques, particularly for non-structural systems, to facilitate rapid repairs and replacements. Both laboratory and in-situ testing of these new materials and building systems should be conducted.

Sattar, Siamak; Segura, Christopher; Johnson, Katherine; McAllister, Therese; McCabe, Steven.
"Building Design Considerations to Support Immediate Occupancy Performance Objectives."
Paper presented at 12th Canadian Conference on Earthquake Engineering, Quebec City, Canada. June 17, 2019 - June 20, 2019.

***Smart buildings*:** Smart buildings that integrate a network of sensors that communicate the condition of a building's various systems can assist in making decisions for evacuation and reoccupation, and can expedite the post-hazard repair process. Research for such smart buildings is needed at both the material level and system level. At the material level, research should prioritize the development of technologies that can communicate material damage and repair needs, as well as materials that can self-repair or adapt to changing conditions. System level research should include the development of new sensors and technologies that can communicate building functionality and identify the systems that are hindering functionality.

***Damage-tolerant, rapidly-repairable building systems*:** An important advancement for the implementation of IO objectives will be the development of new materials and technologies that are either capable of sustaining minimal damage or being rapidly repaired in the aftermath of a hazard event. Material level research is needed to develop new damage-resistant materials (e.g., mold-resistant materials) and increase the accessibility of such materials to the building industry. System level research needs include developing new low-damage technologies, developing systems in which damage is concentrated to a small number of easily-repairable components, introducing new rapid-repair methods to minimize post-hazard downtime, and developing new damage-resistant building envelopes.

***Adoption of new materials and technologies*:** The use of new materials and technologies in the construction industry has been hindered due to a variety of issues including high initial costs, lack of workforce expertise, lack of adequate design guidelines, and liability concerns. Research is needed to better understand how the building industry responds to advances in materials and technologies and what steps are needed to encourage their adoption. Design guidelines and standards should be adapted to provide guidance on new materials and technologies. Additionally, the U.S. should study international communities that have successfully adopted these new materials and technologies in the construction industry.

Maintenance, Repair, and Retrofit Methods

Damage and degradation to a building can reduce the building's structural capacity and ultimately its ability to meet its intended safety and functionality objectives. To improve and optimize building maintenance, research is needed to quantify the impacts of environmental factors on the degradation of building materials and how such degradation affects the performance of building components. The development of low cost, rapid repair technologies is needed to enable timely repair of damaged building components such that buildings designed for IO are able to meet their functionality goals. The development of effective retrofit technologies for existing buildings is also needed to enable the implementation of IO performance objectives. Additionally, research is needed to understand the decision-making processes that influence whether a building owner will choose to invest in maintenance, repair, and retrofit. Recommended research and implementation activities to achieve these goals are summarized below.

***Inventory of existing building stock*:** Data identifying the physical systems and conditions of a community's buildings are needed to inform a number of different studies. Data collection needs include architectural drawings, land use, maintenance and repair records, and information about any building modifications. This information is needed in order to prioritize buildings that might require retrofits or repairs prior to and following a hazard event. In addition, there is a need to collect data about the costs and methods used for repairs and retrofits so that communities and building-owners can make informed decisions to undertake these efforts.

***Decreasing cost and improving methods to repair and retrofit buildings*:** Identifying ways to decrease the costs of repairs and retrofits, as well as ways to expedite repair and retrofit processes, can enable more rapid implementation of IO objectives. Lower repair costs may come as a result of more affordable materials and the development of new rapid repair techniques. In new buildings, both direct and indirect repair costs may be reduced by integrating predefined components designed to sustain the majority of damage and which are easily accessed and repaired. In addition, smart materials that communicate repair needs or are able to self-repair could aid maintenance and repair processes.

***Behavioral research*:** Behavioral research is needed to understand current decision-making processes concerning how building owners choose to invest in maintenance and repair. This research should prioritize understanding when a building owner chooses to upgrade a building, and for what purposes. In addition, research is needed to understand the extent to which building owners value retrofits and repairs to improve hazard resilience.

***Understanding and enhancing repair effectiveness*:** Research is needed to identify and evaluate repair and retrofit techniques for various building types, including the repair or retrofit of building envelopes and the nonstructural systems of the buildings. The effectiveness of these repair and retrofit methods in restoring and enhancing the strength, safety, and functionality of a building should be studied.

Sattar, Siamak; Segura, Christopher; Johnson, Katherine; McAllister, Therese; McCabe, Steven.
"Building Design Considerations to Support Immediate Occupancy Performance Objectives."
Paper presented at 12th Canadian Conference on Earthquake Engineering, Quebec City, Canada. June 17, 2019 - June 20, 2019.

***Understanding resources needed for repairs and retrofits*:** Detailed analytical studies are needed to investigate the financial and social impacts associated with repairing various building types damaged under a variety of hazards scenarios. The costs associated with retrofitting these buildings prior to the occurrence of various hazard scenarios should also be studied. A comparison of the direct mitigation costs to the potential socioeconomic costs over a building's lifecycle can help inform decision makers during the building design, repair, and retrofit decision-making processes.

***Improving availability of tools, parts, and labor*:** In the aftermath of a hazard event causing widespread damage, demand is often high for skilled labor, specialty equipment, and custom manufactured parts. This demand is often unable to be met immediately following a hazard event, leading to long recovery timeframe for impacted communities. Understanding and improving the availability of tools, parts, and skilled labor for recovery could help reduce the amount of time required for building repairs, decreasing the disruption on the community, building owners, and occupants.

***Methods to strategically implement retrofits*:** Due to the high costs associated with retrofitting a community's vulnerable building stock, a retrofit prioritization scheme may be necessary to determine which buildings should be retrofitted first, and how those retrofits should take place to minimize interruptions to building owners and occupants. Building rating systems should be explored to determine if any existing or new systems could be implemented to prioritize which buildings should be repaired and retrofitted.

Developing periodic inspection protocols: Buildings designed to immediate occupancy performance objectives should be maintained and inspected periodically to ensure that the buildings continue to meet IO objectives. This is not part of current practice, and therefore would likely require research to determine cost-effective techniques for implementation.

Monitoring and Assessment

The state of an IO building should be monitored periodically to determine how aging, environmental factors, and hazards affect the building's ability to meet IO performance objectives. Moreover, assessing the building's performance after a natural hazard event is essential to evaluate whether the building is safe to occupy and to determine the post-hazard functionality level of the building. The main challenge in monitoring building performance is developing new cost-effective monitoring techniques. Timely assessment of buildings after a hazard is also a key challenge, as it needs to be completed prior to reoccupying the building and returning to function. Recommended research and implementation activities to achieve these goals are summarized below.

Technology and sensors to assess building performance: Research is needed to identify effective technologies and methods for conducting rapid assessments of building condition. This effort includes developing cost-effective sensors and built-in monitoring systems for collecting data, as well as developing performance assessment procedures to analyze the collected data and identify the observed damage. One of the main needs in improving data collection technologies is the development of instrumentation that can monitor the response of nonstructural systems. Research is also needed to develop sensors that can monitor the degradation of material properties due to environmental impacts.

New data collection methods: Research is needed to investigate the use of new data collection methods, such as crowdsourcing, social media, and use of drones for collecting information for monitoring and assessment of buildings after natural hazards.

Linking damage measurement to the performance assessment: Data collected from sensors or similar technologies need to be processed either by an engineer or through an automated process to quantify building damage after a hazard event. Research is needed to relate recorded data to observed building damage, post-hazard functionality, and recovery time. This research effort includes benchmarking/validating the linkage between the collected data and observed damage using available data from instrumented buildings. This research is an important step toward developing a remote assessment system for buildings, where collected data is automatically sent to and analyzed by an assessment tool to identify the extent of damage. Moreover, research is needed to help develop methods to use recorded data to assist with the decision-making processes after a hazard event.

Sattar, Siamak; Segura, Christopher; Johnson, Katherine; McAllister, Therese; McCabe, Steven.
"Building Design Considerations to Support Immediate Occupancy Performance Objectives."
Paper presented at 12th Canadian Conference on Earthquake Engineering, Quebec City, Canada. June 17, 2019 - June 20, 2019.

Improving inspection techniques: Currently, post-event evaluation of a building is conducted primarily through visual inspection, which may be adequate to assess collapse likelihood in a general sense but may not be sufficient to identify the safety and functionality of a building designed for IO. Research is needed to develop new post-hazard evaluation criteria, specifically for IO buildings, and to develop an IO building tagging process. One of the key issues is to identify who is responsible for tagging of IO buildings and what improvements need to be made to the current tagging process to shorten the inspection period.

Developing guidelines/protocols for inspection of the buildings for immediate occupancy: To quantify the preserved level of occupancy and functionality of a building throughout its lifecycle, guidelines and protocols on implementing inspection requirements and methods should be developed.

CONCLUSIONS

This paper articulates research needs and implementation activities concerning the design of individual buildings that are necessary to support development of IO performance objectives. These include: defining acceptable level of functionality for IO buildings; quantifying damage levels that are appropriate for the distinct functionality levels; developing new building materials and construction techniques for IO buildings; maintaining and retrofitting buildings to meet IO objectives; and monitoring and assessing the state of a building throughout its lifecycle. Unlike current building code objectives, the new IO performance objective prioritizes maintaining an appropriate level of building functionality, including retaining functionality, following a hazard event. Therefore, the research needs and activities described in this paper will require a fundamental change to the current state of practice. In addition, current practice for retrofitting and monitoring existing buildings will require substantial change. Research under the topics identified in this paper could lead to the development of new design guidelines and analytical tools to assist developers, architects, engineers, and researchers in designing and assessing buildings to achieve IO performance objectives. Development and adoption of guidelines and tools to implement IO objectives would advance current standards of practice and lead to buildings that are more resilient to natural hazards, providing a greater level of safety and minimizing disruptions for building occupants.

While this paper focuses on the technical aspect of design of individual buildings to meet IO objectives, the successful development and adoption of IO objective requires a broader perspective that considers the interactions between individual buildings and the surrounding community, the social and economic impacts of IO buildings on different stakeholders, and methods to garner public support to ensure successful adoption of IO objectives. These diverse issues demand multidisciplinary perspectives and engagement from all levels of society.

ACKNOWLEDGMENTS

This project was completed with assistance from personnel at the Science and Technology Policy Institute (STPI) of the Institute for Defense Analyses (IDA), located in Alexandria, VA. The authors also gratefully acknowledge contributions from a number of groups: the steering committee members Mary Comerio, Gregory Deierlein, Susan Dowty, John Gillengerten, James Harris, William Hirano, Laurie Johnson, Timothy Reinhold, and James Rossberg; workshop participants; and reviewers of the NIST report.

REFERENCES

[1] U.S. Senate; S. Rep. No. 114-239, at Disaster Resilient Buildings. (2016), Retrieved from GPO's Federal Digital System: https://www.gpo.gov/fdsys/pkg/CRPT-114srpt239/html/CRPT-114srpt239.htm.

[2] Sattar, S., McAllister, T., Johnson, K., Clavin, C., Segura, C., McCabe, S., Fung, J., Abrahams, L., Sylak-Glassman, E., Levitan, M., Harrison, K., Harris, J. (2018) "Research Needs to Support Immediate Occupancy Building Performance Following Natural Hazards," National Institute of Standards and Technology, NIST SP-1224, Gaithersburg, MD, U.S.A.

Sattar, Siamak; Segura, Christopher; Johnson, Katherine; McAllister, Therese; McCabe, Steven.
"Building Design Considerations to Support Immediate Occupancy Performance Objectives."
Paper presented at 12th Canadian Conference on Earthquake Engineering, Quebec City, Canada. June 17, 2019 - June 20, 2019.

A Collaborative Work Cell Testbed for Industrial Wireless Communications — The Baseline Design

Yongkang Liu*, Richard Candell†, Mohamed Kashef*, Karl Montgomery†
*Advanced Network Technologies Division †Intelligent Systems Division
National Institute of Standards and Technology, Gaithersburg, Maryland, USA
Email: {yongkang.liu, richard.candell, mohamed.hany, karl.montgomery}@nist.gov

***Abstract*—A work cell is an essential industrial environment for testing wireless communication techniques in factory automation processes. A new testbed was recently designed and developed to facilitate such studies in work cells by replicating various data flows in an emulated production environment. In this paper, the testbed's baseline design is presented which characterizes deterministic and reliable communication needs between work cell components in a typical machine tending application. Special design issues are discussed regarding safety measures in collaborative robotic operations and network synchronization among distributed machines. Measurement plans in the hardwired baseline are also introduced along with further wireless extensions. The testbed can serve as a representative cyber-physical system (CPS) model to verify industrial wireless techniques in support of mission-critical data transmissions.**

***Index Terms*—industrial wireless communications, industrial wireless networks, industrial wireless testbed, factory automation processes, testbed design.**

I. Introduction

Industrial communication networks leverage operational technology (OT) insights/decisions in recent Industry 4.0 and Smart Manufacturing initiatives through mission-critical data sharing between field instruments and factory automation controllers [3], [4]. Compared with hardwired connections, wireless links have unique advantages in connecting field sensors and actuators with reduced cabling cost and natural support of mobility [1]. A number of industrial wireless solutions have been proposed for improved production efficiency, asset health, and workplace safety [2]. However, they need to prove the full support on agile plant operations with trusted transmission timeliness and reliability before being adopted on the factory floor. Evaluating the capabilities of various and diverse wireless technologies has turned out to be a challenging but essential task to promote industrial wireless applications [5], [6].

An evaluation platform provides necessary details of performance requirements and operation specifications in typical industrial wireless use cases, e.g., the plant layout, process workflow, wireless channel model, and data traffic pattern. It plays an important role in verifying wireless network design and comparing the performance of different wireless technologies. Such modeling efforts have been taken both at the macroscopic level, e.g., on spatial statistics of the node density and traffic load on the factory floor, and at the microscopic level, e.g., on the latency and interference in individual transceiver pairs. Based on these models, system verification methods using co-simulation platforms [7], [8], [9], [10], hardware-in-the-loop (HIL) experiments [11], and testbeds [12] become popular in studying the unique industrial environments and service characteristics.

A few new challenges have emerged which need a further investigation when modeling various plant factors. First, most evaluation workflows are one-way, i.e., describing the impact from industrial environments and operations onto wireless transmissions. Since industrial systems, as complex cyber-physical systems (CPS), are featured with the interplay between industrial processes and data networks, the model is expected to represent interactive connections between OT and information technology (IT) systems. Second, machine-to-machine (M2M) communications carry and distribute data in a vastly different way compared to the conventional Internet data. The model also needs to characterize and verify various traffic patterns, both empirical and statistical. Last but not least, current models are usually built upon snapshots of existing industrial practices which only capture environments and activities of the status quo for the network design and optimization. As CPS innovations have been evolving in emerging industrial use cases, the new model has to be more flexible in compliance with both short- and long-term network implementations.

Through measurements of process and network activities to finely tune performance requirements on industrial wireless networks, a testbed is being developed at the National Institute of Standards and Technology (NIST). This paper introduces the baseline design which identifies a variety of data needs in the emulated industrial operations and calibrates the performance under hardwired connections before extending to wireless alternatives. Generally, the testbed is featured with three aspects of innovations.

First, the testbed picks a work cell as the target model which is at the "right" size to capture essential data traffic patterns between industrial devices in a manufacturing cycle. By inspecting both internal module coordination and external interactions with upper-level management systems, the proposed work cell testbed serves as a good reference to verify industrial wireless networks in supporting efficient manufacturing operations.

Second, the testbed is specialized in emulating collaborative operation scenarios various machining tools working with industrial robots. Cooperations between machines and their

Liu, Yongkang; Candell, Richard; Hany, Mohamed; Montgomery, Karl.
"A Collaborative Work Cell Testbed for Industrial Wireless Communications - The Baseline Design."
Paper presented at 2019 IEEE 28th International Symposium on Industrial Electronics, Vancouver, Canada. June 12, 2019 - June 14, 2019.

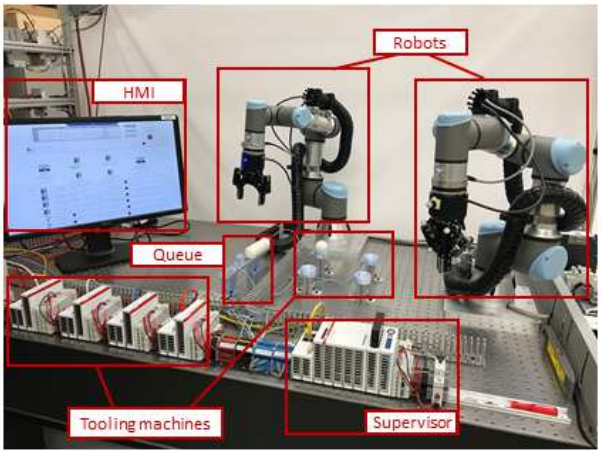


Fig. 1: Collaborative work cell testbed

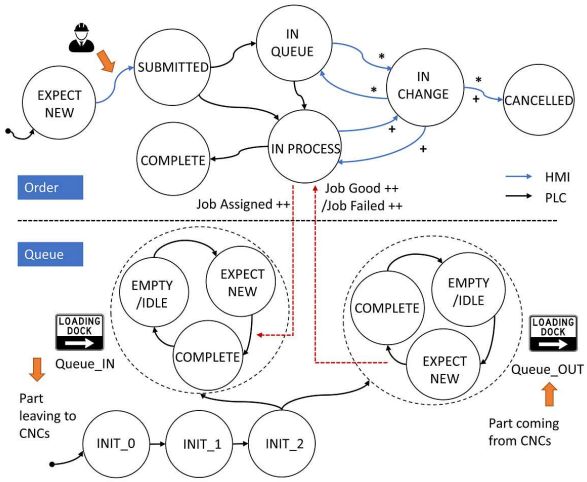


Fig. 2: State machines of the order and queue modules

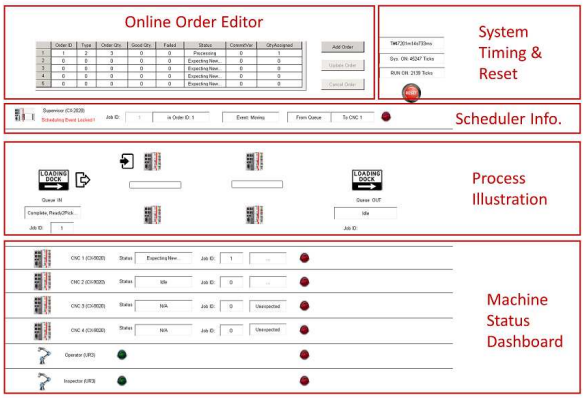


Fig. 3: Snapshot of the testbed human-machine interface

robotic partners are managed by the work cell supervisor through customized data flows, e.g., the context information, size, and frequency, following industrial specifications.

Third, the testbed provides rich work cell footprints in production operations which facilitates network measurements and evaluation. Compared with previous modeling efforts which simply treated individual work cells as buffers of working parts/orders [10], the testbed further characterizes data flows within and beyond work cells to fully represent data features in complex industrial scenarios.

The remainder of the paper is organized as follows. The system architecture is introduced along with brief discussions on the emulated production processes in Section II. Details about the design of machine emulators are presented in Section V. The network synchronization issues and safety-related operations are discussed in Section VI and Section IV, respectively. The ongoing measurement and wireless extensions are introduced in Section VII. Concluding remarks are given in Section VIII.

II. Overview of Testbed Design

A. Work Cell Components

As shown in Fig. 1, the testbed emulates a generic work cell in the manufacturing factory which consists of multiple components including a supervisory control unit, machines, interstage buffers, robots, and human workers.

1) Supervisor: The supervisory control unit, or supervisor, manages its work cell by monitoring the whole production process, scheduling production based on incoming orders, and coordinating inter-node actions. Meanwhile, it also serves as the agent on behalf of the entire work cell to communicate with the upper-level managing systems in the factory, such as supervisory control and data acquisition systems (SCADA) and manufacturing execution systems (MES). A programmable logic controller (PLC) usually plays the supervisor's role in the work cell. In the testbed, we use a Beckhoff CX2020 PLC as the work cell supervisor which is equipped with various communication interfaces for internal and external information exchanges [13].

2) Interstage Queue: The interstage queue is comprised of two loading zones in the work cell, i.e., the input (Queue_IN) and output (Queue_OUT) buffers, which serve as the start and end points for a single job, respectively. As shown in Fig. 2, the input buffer accommodates the incoming raw parts into the work cell and the output buffer collects the finished parts, either good or failed. The supervisor detects the arrival/departure events in the buffers with proximity sensors, one for each, and updates the order status accordingly.

3) Machines: Four computer numerical control (CNC) machines are considered in the testbed, whose behaviors in the machine tooling and communications are characterized by emulation models. Each CNC machine consists of a PLC, a part holder, and a proximity sensor. The PLC mimics state transitions of the CNC machine in its tooling cycle and exchanges the machine status and job information with the supervisor. The part holder represents the machine's working zone where the proximity sensor is used to monitor the part arrival/departure. The PLC connects the sensor to its digital input/output (I/O) module and samples the input signal. Four

Liu, Yongkang; Candell, Richard; Hany, Mohamed; Montgomery, Karl.
"A Collaborative Work Cell Testbed for Industrial Wireless Communications - The Baseline Design."
Paper presented at 2019 IEEE 28th International Symposium on Industrial Electronics, Vancouver, Canada. June 12, 2019 - June 14, 2019.

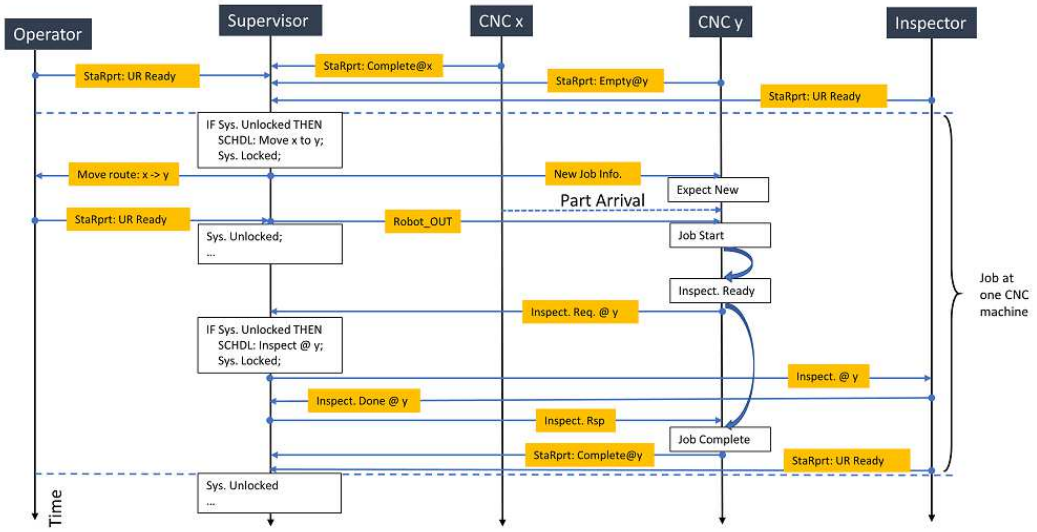


Fig. 4: Timeline illustration of communication messages in an intermediate tooling procedure

Beckhoff CX9020 PLC are used as the emulators along with the propriety I/O modules [14]. Details of the emulator design are discussed in Section V.

4) Robotic Laborers: Two UR3 robots are used in the testbed [15]. Each robot has six degrees of freedom (6 DoF) and is equipped with a gripper and a 6 DoF force torque sensor [16], [17]. Robots mainly communicate with the supervisor to receive actuation commands and report their status. Based on each robot's role in work cell operations, UR3 programs perform motion commands such as waypoint selection and trajectory planning.

5) Human Workers: A collaborative work cell may be operated by human workers or not. In the testbed, human workers can remotely monitor and interact with the automated production process, such as placing orders and stopping/resetting the production, through a human-machine interface (HMI) as shown in Fig. 3. The real-time status information as displayed by HMI is updated through the supervisor and collected from distributed components.

B. Baseline Use Case: Machine Tending

The baseline design studies a machine tending use case. Jobs are assigned to the work cell in batches through the HMI as shown in Fig. 2. Each batch, namely an order, contains a number of jobs/parts of the same type with a specific tool path, i.e., a sequence of moves operated at one or more machines. The two robots play different roles in the production: one as the operator (OPT) and another as the inspector (INS). OPT is in charge of transporting parts between job stops. A job stop refers to the working zone of a machine or the input/output loading zone. The INS robot checks the part quality after each tooling step and reports the inspection result to the supervisor. Based on the result, the supervisor then orders the operator to either move the part to the next stop along the path (if it passed the check) or drop it to Queue_OUT with a defect mark (if it failed). The inspection result is simulated at the inspector by a random variable associated with the emulated tooling operation.

TABLE I: Exemplary specifications of data flows between work cell components

Link	Data	Update Rate	Size (Bytes)	Protocol
Supervisor -CNC	Status report	1 Hz - 100 Hz	10s	ADS
	Safety	100+ Hz	10s	ADS
	Inspection request/response	On-demand	10s	ADS
CNC-CNC	Motion control	1000 Hz	A few	ADS
Supervisor -Robot	Actuation	1 Hz - 50 Hz	A few	Modbus
	Safety	125 Hz	A few	Modbus
Robot -Peripheral	6 axis force and torque sensor	100 Hz - 500 Hz	100s	TCP/IP
Supervisor -External	HMI	10 Hz - 50 Hz	100s	ADS
	IoT	>1 Hz	10s - 100s	MQTT

C. Work Cell Communications

The topology of the work cell network is centered around the supervisor which acts as the information hub and gateway for both internal and external data flows. Fig. 4 illustrates messages that are transported between work cell components in a job move. Connections within and beyond the work cell are managed by different communication protocols. Among them, the inter-PLC links are carrying transmission control protocol/Internet protocol (TCP/IP) based TwinCAT Automation Device Specification (ADS) messages [18]. ADS is a medium-independent protocol for the communication between Beckhoff's TwinCAT devices. The supervisor communicates with robots through Modbus which allows the data exchange between heterogeneous industrial appliances in the shared registers at the supervisor.

Liu, Yongkang; Candell, Richard; Hany, Mohamed; Montgomery, Karl.
"A Collaborative Work Cell Testbed for Industrial Wireless Communications - The Baseline Design."
Paper presented at 2019 IEEE 28th International Symposium on Industrial Electronics, Vancouver, Canada. June 12, 2019 - June 14, 2019.

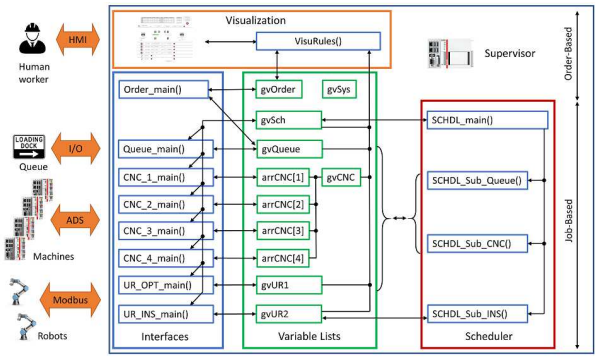


Fig. 5: Architecture of work cell supervisor functions implemented in the PLC

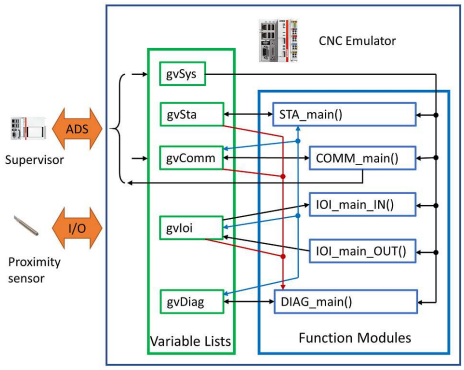


Fig. 6: Architecture of the CNC emulator

Generally, the data exchange in a work cell is determined by the associated production operations. For process variables (PV) regarding the production efficiency, the supervisor needs to collect the updates from remote machines to estimate the loads of individual stations and ensure the quality. For the ones with the asset health, the supervisor uses them to schedule the maintenance downtime and estimate the cost. To coordinate the collaborative operations in the work cell, the real-time status of a machine should be made known to its partners so that the synchronous operation can mitigate errors and improve the quality. Besides routine exchanges, part of the CNC machine data is state-related, i.e., data are transmitted according to the current state in which the machine stays. Table I summarizes the emulated data flows in the testbed.

III. Design of Work Cell Supervisor

As shown in Fig. 5, the supervisor of the testbed consists of four main function blocks in its architecture: the scheduler (SCHDL), interfaces, visualization, and global variable lists (GVL). Specifically, the scheduler is in charge of assigning production jobs to machines and robots. The interface block handles the communications with various CNC machines and UR3 robots in the work cell. Meanwhile, it also updates the order and queue information in emulation experiments. Using the PLC's visualization library, the visualization block controls the testbed's HMI. The supervisor's first three blocks are implemented as the PLC function modules which maintain their own status locally. The system-wide data sharing between function modules takes place in GVL. System variables associated with a specific function or work cell component are managed in the GVL named after it, e.g., "gvCNC" contains four array objects each of which stores the relevant information of a CNC machine in the work cell.

From the perspective of information processing, functions and data memories in the supervisor can be divided into two planes: order-based and job-based. Order-based functions deal with incoming orders, update the order status based on real-time production results, and maintain the inventory. On the other hand, job-based functions mainly work on the associated work cell components and coordinate their production activities following the schedule. Such a modular design allows the supervisor to easily adapt to the composition of a specific work cell and utilize the state-of-the-art techniques to leverage individual functions.

IV. Safety-aware Scheduling and Operations

In the baseline, the testbed considers production activities without physical human contact where human workers stay in the remote safety zone and interact with the process through HMI. Major safety concerns include collision risks between robots and the interruption of machining when a robot hits the running machine. Therefore, the testbed is designed with multiple safety approaches to eliminate possible risks to protect the asset.

First, the supervisor sets a safety flag in its scheduler to indicate if there is an active robot moving in the work cell. The scheduler only assigns at most one robot to be actively operating. Once the flag is set, the locked scheduler would not assign a new job to another robot so that collisions are avoided.

Second, the active robot will keep notifying the contacted machine(s) in the current job so that the machine would not start to process the part until the robot returns to the safety zone. As shown in Fig. 4, the "Robot_OUT" message indicates the clearance of the contact.

Besides, an additional logic check on the waypoint information is performed at the robot to verify the fetched instruction through Modbus. Meanwhile, the supervisor will clear the waypoint information set in the registers right after the robot confirms the reception. In this way, it prevents the robot from repeating out-of-date operations in case that the new waypoint information is lost in the transmission. Initial experiments confirm that introducing such an approach allows error-free operations through very light supervisor-robot Modbus communications as low as 1 Hz.

V. Machine Tool Emulation

The testbed is aimed to evaluate the mutual impact between data transmissions and the work cell performance. Therefore, the CNC emulator is mainly focused on mimicking the machine's behaviors with time dependent and statistical performance features, such as the production efficiency, error

Liu, Yongkang; Candell, Richard; Hany, Mohamed; Montgomery, Karl.
"A Collaborative Work Cell Testbed for Industrial Wireless Communications - The Baseline Design."
Paper presented at 2019 IEEE 28th International Symposium on Industrial Electronics, Vancouver, Canada. June 12, 2019 - June 14, 2019.

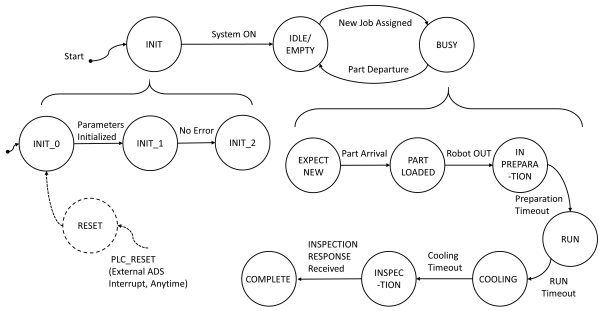


Fig. 7: CNC state machine

and downtime distributions, and part defects. Meanwhile, the emulators also help shape the work cell data traffic with their periodic status updates and on-demand messages during the production. Following the similar modular design as the supervisor, the CNC emulator also defines its function modules and GVL in the implemented PLC. Specifically, function modules include the state machine (STA), communications (COMM), I/O module interfaces (IOI)[1], and diagnostics (DIAG). The associated system variables shared in between are organized in GVL, e.g., "gvSta" maintains variables related with the state machine and "gvSys" contains the system-wide information such as the machine's identification (ID) and network address.

To fully capture the operational and communication activities of a machine tool, the CNC emulator conducts state-dependent operations and communications characterized by the state machine as shown in Fig. 7.

The state machine is defined in the STA function module which contains three main states: initialization (INIT), idle, and busy. Each main state can contain a few substates which characterize further details of operations. INIT along with its substates facilitate the synchronization among distributed nodes whose design will be discussed with more detail in Section VI. The substates of the busy state represent a series of machine operations regarding a single job. The dwelling time in each (sub-)state can be either timed according to the machine's specification, e.g., the approximate G-code execution time and material removal rate, or determined by external events that trigger state transitions, e.g., a notification message. The randomness can also be introduced based on statistical machine/production models. Examples of randomness components in the models include: 1) the time of a tooling procedure; 2) time varying energy consumption in different states, e.g., power variations in material-drilling processes; 3) tool life estimation; 4) part defect rate; 5) measurement drift between calibrations; and 6) safety related events, e.g., unexpected interrupts due to object intrusion. Using empirical models and measurement data, we can model the above performance metrics statistically and regenerate the state-related traffic for the studied machine.

Therefore, the machines emulated in the testbed can be programmed to highlight the details of real practices to study the network impact on the work cell performance. PV can be modeled in the testbed focusing on different topics such as 1) the production (task) efficiency, e.g., the execution time, material removal rate, energy consumption, and part defect rate; 2) asset health, e.g., the tool life time, failure probability, and downtime schedules for calibration and maintenance; and 3) work cell collaboration, e.g., the clock drift, coordination precision, and safety. Besides checking the network support on routine data transmissions as scheduled, the testbed is particularly useful for testing the network performance in extreme cases with rare occurrences. The machine emulator can produce the traffic in the special use cases, such as the recovery from unexpected overload events or in emergency cases, and repeat it for comparative studies.

The quality of the "product" is also virtually rendered in the testbed. The result of each single part after a machining process is randomly generated following the statistical model to mimic the defect rate in a real machine. The inspector is in charge of generating the result and returning it to the supervisor for scheduling the next move. According to the study of the quality and quantity relationship in production systems [19], [20], part failures have both independent and dependent causes. The independent failure follows a Bernoulli distribution with the uncertainty of temporal independence. On the other hand, the dependent types of failures, which are often referred to as "persistent" or "systematic" ones, are those caused by tool failures, such as the broken drill or clog in the painting tube. In such cases, the failure of product is highly related with the asset failure rate. Since both types of failures are decoupled by their nature, the testbed carries the failures of the product as well as the ones related with assets to emulate the occurrences of various failures across time. The delivery delay or loss in communication links also affect the performance of operations and safety measures.

VI. Synchronization of Networked Components

Since work cell components are collaboratively working in the production, the testbed implements multiple approaches to coordinate these distributed nodes.

First, we develop a phased initialization process at the beginning of each experiment. The testbed initialization includes three steps:

INIT_0: Parameter initialization/reset;
INIT_1: Logic error check and confirmation; and
INIT_2: Loading ready-to-go state.

The supervisor keeps the pace by triggering the state transition only after all components have met the state-specific conditions. Meanwhile, the testbed also supports the online reset through the HMI as shown in Fig. 3. Once the reset button is clicked, the supervisor will send the reset commands to individual nodes and direct them to restart from INIT_0.

Besides signaling procedures, the testbed also introduces global clock synchronization throughout the work cell. In the work cell, a Meinberg Lantime M900 time server provides the IEEE 1588 precision time protocol (PTP) synchronization service as the grandmaster [21]. The supervisor PLC is

[1]The IOI functions are further grouped into IOI IN and IOI OUT, respectively.

Liu, Yongkang; Candell, Richard; Hany, Mohamed; Montgomery, Karl.
"A Collaborative Work Cell Testbed for Industrial Wireless Communications - The Baseline Design."
Paper presented at 2019 IEEE 28th International Symposium on Industrial Electronics, Vancouver, Canada. June 12, 2019 - June 14, 2019.

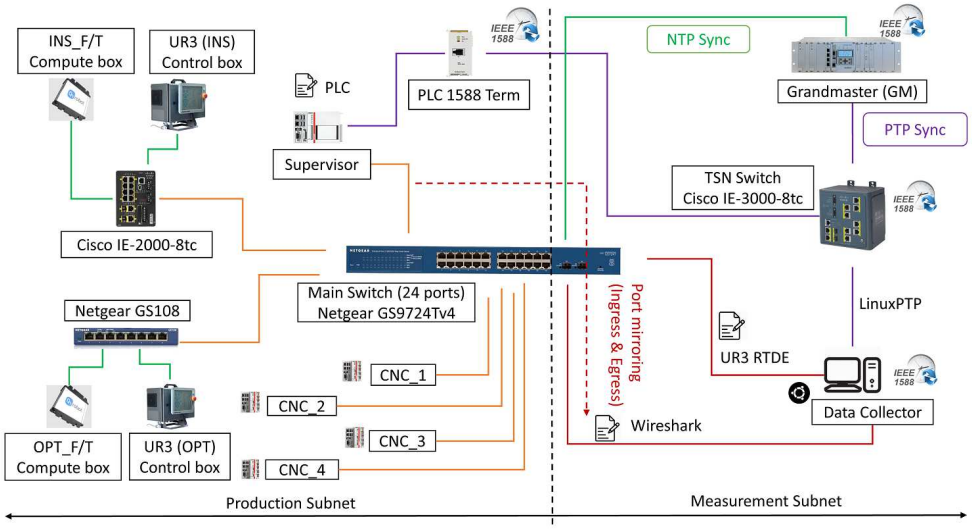


Fig. 8: Illustration of network topology in baseline measurements. Green and orange lines indicate connections for production data and process-related traffics where the orange ones are candidates to be replaced by wireless alternatives in the next phase. Purple lines are used for the PTP synchronization purpose; red lines carry measurement data.

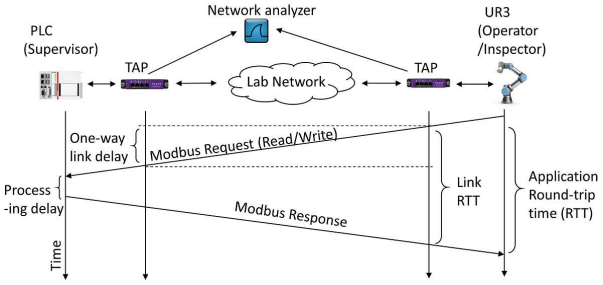


Fig. 9: Illustration of using network TAP devices in the link level delay measurement

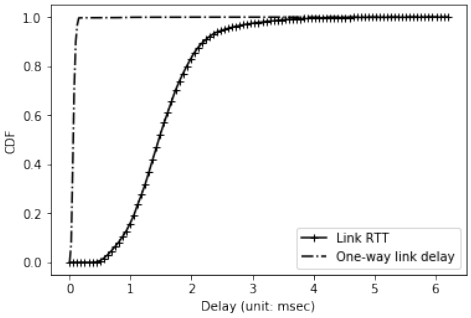


Fig. 10: Result of the link level delay measurement of Modbus transactions in the baseline network

equipped with a Beckhoff IEEE 1588 terminal to synchronize its local clock with the time server [22]. Measurement devices also run the LinuxPTP software to render time stamps in collecting the real time status of UR3 robots and network traffic captures [23].

VII. Measurements and Wireless Extensions

A. Testbed Measurements

System and network measurements are performed in the testbed which employs various performance metrics regarding the production efficiency, product quality, and network utility in highly discrete manufacturing processes [24]. An illustration of the network topology used in baseline measurements is shown in Fig. 8. The main observation point for network traffic is set at the supervisor as the testbed takes a centralized topology. Table I also indicates that the majority of data flows assumed in the work cell operations are associated with the supervisor. Therefore, data that are routed from/to the supervisor are collected. Specifically, all work cell components are connected to an industrial-grade switch whose ports are further separated into production operation and measurement uses. Utilizing the switch's "port mirroring" function, we copy and forward the data from the supervisor's operation port to the measurement port where a computer collects the data with network packet analyzers, e.g., WireShark. Data collected in individual experiments will be stored for future analysis and modeling.

As part of the proof of concept, we introduce the network test access point (TAP) devices in the link-level measurements to study the impact of link-level transmissions on the work cell performance, e.g., packet losses of mission-critical PV updates. As shown in Fig. 9, we use two TAP devices to collect data copies at both ends of a Modbus link between the supervisor and the UR3 robot and employ Using Python's Scapy library to obtain link delay statistics [26]. Fig. 10 presents the cumulative distribution functions (CDFs): one for

Liu, Yongkang; Candell, Richard; Hany, Mohamed; Montgomery, Karl.
"A Collaborative Work Cell Testbed for Industrial Wireless Communications - The Baseline Design."
Paper presented at 2019 IEEE 28th International Symposium on Industrial Electronics, Vancouver, Canada. June 12, 2019 - June 14, 2019.

the round-trip time (RTT) in Modbus transactions between Supervisor and UR3 operator, and another for the one-way link delay. The average values of the link RTT and one-way link delay are 1.528 msec and 0.0627 msec, respectively, in a 3-hop Ethernet path as shown in Fig. 8. A longer delay would be expected along with link failures in lossy wireless channels.

B. Wireless Extensions

Based on the baseline design, wireless extensions are also underway. As each network node is equipped with Ethernet adapter(s), hardwired connections between work cell components can be replaced by wireless links if the Ethernet-wireless adapters are used. Currently, we are working with industrial partners to verify the wireless solution using wireless local area network (WLAN) radios. To reduce the conversion delay between Ethernet packets and WLAN packets, the Ethernet-WLAN conversion takes place in the link layer (Layer 2 forwarding) where both Ethernet packets and wireless packets share the same network address of the node. Channel emulation is also considered in the testbed evaluation to mimic the channel response in real factory radio environments [25]. The time sensitive networking study over wireless links is another target in this project. The synchronized clocks of industrial equipment facilitate collaborative operations, e.g., in coordinated robot movements, and leverage the management of orthogonal time-frequency radio resources.

VIII. Conclusions

In this paper, we have presented a work cell testbed and explained design details for both hardware and software implementations. In addition, measurement techniques and the applicability of wireless links to the design have also been discussed. The testbed is aimed to serve as an evaluation platform for verifying the performance of different wireless technologies in support of deterministic and reliable industrial communications. As an ongoing effort, the current version is built as a baseline with hardwired Ethernet connections between individual components. In future work, we will introduce wireless links and evaluate their performance in harsh industrial radio environments. The future progress and measurement data will be released in the NIST public domain repository as a reference for modeling efforts and comparative studies on industrial wireless technologies [27].

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] L. L. Bello, J. Åkerberg, M. Gidlund and E. Uhlemann, "Guest Editorial Special Section on New Perspectives on Wireless Communications in Automation: From Industrial Monitoring and Control to Cyber-Physical Systems," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1393-1397, June 2017.

[2] V. K. L. Huang, Z. Pang, C. J. A. Chen, and K. F. Tsang, "New Trends in the Practical Deployment of Industrial Wireless: From Noncritical to Critical Use Cases," *IEEE Ind. Electron. Mag.*, vol. 12, no. 2, pp. 50–58, 2018.

[3] H. Kagermann, W. Wahlster, and J. Helbig, "Recommendations for Implementing the Strategic Initiative Industrie 4.0", Industrie 4.0 Working Group, 2013.

[4] A. B. Feeney, S. Frechette and V. Srinivasan, "Cyber-Physical Systems Engineering for Manufacturing", in *Industrial Internet of Things - Cybermanufacturing Systems*, S. Jeschke et al., Eds. Springer Nature, 2017, pp. 81–110.

[5] J. Ansari et al., "Demo: a realistic use-case for wireless industrial automation and control," in *Proc. NetSys 2017*, Gottingen, 2017, pp. 1-2.

[6] Y. Liu, R. Candell, M. Kashef, and L. Benmohamed, "Dimensioning Wireless Use Cases in Industrial Internet of Things," in *Proc. IEEE WFCS'18*, Imperia, Italy, Jun. 2018.

[7] E. Galli, G. Cavarretta and S. Tucci, "HLA-OMNET++: an HLA compliant network simulation," in *Proc. DS-RT'08*, pp. 319-321, 2008.

[8] H. Neema, et. al. "Model-Based Integration Platform for FMI Co-Simulation and Heterogeneous Simulations of Cyber-Physical Systems," in *Proc. 10th International Modelica Conference*, 2014.

[9] Y. Liu, R. Candell, K. Lee, and N. Moayeri, "A Simulation Framework for Industrial Wireless Networks and Process Control Systems," in *Proc. IEEE WFCS'16*, Aveiro, Portugal, May 2016.

[10] J. Geng et al., "Model-based cosimulation for industrial wireless networks," in *Proc. WFCS 2018*, Imperia, 2018, pp. 1-10.

[11] J. Kölsch, C. Heinz, S. Schumb and C. Grimm, "Hardware-in-the-loop simulation for Internet of Things scenarios," in *Proc. Workshop MSCPES 2018*, Porto, 2018, pp. 1-6.

[12] R. Candell, T.A. Zimmerman, and K.A. Stouffer, "An Industrial Control System Cybersecurity Performance Testbed", NISTIR 8089, Dec. 2015.

[13] Beckhoff, "CX2020 Basic CPU Module", 2018. [online]. Available: https://www.beckhoff.com/english.asp?embedded_pc/cx2020.htm. [Accessed January 6, 2019].

[14] Beckhoff, "CX9020 Basic CPU Module", 2018. [online]. Available: https://www.beckhoff.com/english.asp?embedded_pc/cx9020.htm. [Accessed January 6, 2019].

[15] Universal Robots, "Universal Robot UR3", 2018. [online]. Available: https://www.universal-robots.com/products/ur3-robot/. [Accessed January 6, 2019].

[16] Robotiq, "2F-85 and 2F-140 Grippers", 2019. [online]. Available: https://robotiq.com/products/2f85-140-adaptive-robot-gripper. [Accessed on January 10, 2019].

[17] OnRobot, "6 axis Force Torque Sensor", 2018. [online]. Available: https://onrobot.com/products/hex-force-torque-sensors/. [Accessed January 6, 2019].

[18] Beckhoff, "ADS Introduction", Beckhoff Information System, 2018. [online]. Available: https://infosys.beckhoff.com/. [Accessed January 6, 2019].

[19] I. C. Schick, S. B. Gershwin, and J. Kim, "Quality/Quantity Modeling and Analysis of Production Lines Subject to Uncertainty, Phase I, Final Report", May 2005. [online]. Available: http://cell1.mit.edu/papers/GM_PhaseI_FinalReport-2005.pdf

[20] J. Kim and S. B. Gershwin, "Integrated quality and quantity modeling of a production line", OR Spectrum, Vol 27, No. 2-3, pp. 287–314, June 2005.

[21] Meinberg, "LANTIME M900/PTP", 2018. [online]. Available: https://www.meinbergglobal.com/english/products/modular-3u-ieee-1588-grandmaster-clock.htm. [Accessed January 6, 2019].

[22] Beckhoff, "IEEE 1588 external synchrnozation interface (EL6688)", Beckhoff Information System, 2018. [online]. Available: https://www.beckhoff.com/english.asp?ethercat/el6688.htm. [Accessed April 25, 2019].

[23] Red Hat, "Chapter 23. Configuration PTP using PTP4L", in *Deployment, Configuration and Administration of Red Hat Enterprise Linux 6*. [online]. Available: https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/6/html/deployment_guide/ch-configuring_ptp_using_ptp4l

[24] R. Candell, K. A. Stouffer, and D. Anand, "A Cybersecurity Testbed for Industrial Control Systems", in *Proc. PCS 2014*, Houston, TX, Oct. 2014.

[25] R. Candell et al., "Industrial Wireless Systems Radio Propagation Measurements", NIST Technical Note 1951, 2017.

[26] Scapy, "Packet crafting for Python2 and Python3", 2019. [online]. Available: https://scapy.net/ [Accessed on April 25, 2019].

[27] NIST, "Wireless Systems for Industrial Environments", 2019. [online]. Available: https://www.nist.gov/programs-projects/wireless-systems-industrial-environments [Accessed on January 7, 2019].

Liu, Yongkang; Candell, Richard; Hany, Mohamed; Montgomery, Karl.
"A Collaborative Work Cell Testbed for Industrial Wireless Communications - The Baseline Design."
Paper presented at 2019 IEEE 28th International Symposium on Industrial Electronics, Vancouver, Canada. June 12, 2019 - June 14, 2019.

CIRP Manufacturing Systems Conference 2019

Toward data-driven production simulation modeling: dispatching rule identification by machine learning techniques

Satoshi Nagahara[a]*, Timothy A. Sprock[b], Moneer M. Helu[b]

[a]*Hitachi, Ltd., Research & Development Group, 292 Yoshida-cho, Totsuka-ku, Yokohama-shi, Kanagawa, 244-0817, Japan*
[b]*National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland, 20899, USA*

* Corresponding author. Tel.: +1-50-3135-3415; fax: +1-50-3135-3412. *E-mail address:* satoshi.nagahara.eb@hitachi.com

Abstract

Production simulation is useful to predict and optimize future production. However, it requires effort and expertise to create accurate simulation models. For instance, operational control rules, such as job sequencing rules, are modeled based on interviews with shop-floor managers and some assumptions since those rules are tacit in general. In this paper, we consider a data-driven approach to model operational control rules. We develop job sequencing rule identification methods that model rules from production data using machine learning techniques. These methods are evaluated based on accuracy and robustness against uncertainty in human decision making using virtual and real production data.

Peer-review under responsibility of the scientific committee of the 52nd CIRP Conference on Manufacturing Systems.

Keywords: Production simulation, dicrete event simulation, dispatching rule, job sequencing rule, learning to rank

1. Introduction

Market needs have diversified over the years, making production systems more complex. For complicated production systems, production simulation, an embodiment of discrete event simulation (DES), is a well-known and powerful tool to evaluate, plan, and control production. In that case, it is required to build simulation models that emulates the real system accurately. Simulation inaccuracy directly affects the performance of simulation-based applications such as decision-support systems, scheduling systems, etc. However, building accurate simulation models is a time-consuming task, requiring both production domain knowledge and simulation expertise [1]. It makes the practical use of production simulation difficult. Therefore, automating production simulation modeling is still a challenging and important issue [2].

A general modeling process of production simulation is described as follows (adapted from [1]). First, characteristics of production systems are investigated mainly by field-work and interviews with shop-floor managers. Through the investigation, general requirements, such as scope and granularity of production simulation, are defined. In the second step, data required for simulation is collected and/or generated. In the third step, simulation models are constructed from the data obtained in the second step. Then, simulation accuracy is evaluated by comparing simulation results to historical production data. Lastly, these steps are iterated until the required accuracy is achieved.

To reduce the effort in this modeling process, the automation of the second step (data collection and generation) is crucially important. Barlas et al. [3] state that collecting and generating simulation data is still one of the most time-consuming tasks and a barrier to applying production simulation in practice. They also summarized existing studies on automatic collection and generation of simulation data, classifying the studies into five categories: intermediary database, PLC programs, developed applications, data interfaces / standards translators, and direct integration.

Simulation data is mainly composed with structural information such as process route, time information such as processing time, and information regarding operational control rules. One idea to automate this step is to utilize data stored in

Nagahara, Satoshi; Sprock, Timothy; Helu, Moneer.
"Toward data-driven production simulation modeling: dispatching rule identification by machine learning techniques."
Paper presented at 52nd CIRP Conference on Manufacturing Systems (CMS 2019), Ljubljana, Slovenia. June 12, 2019 - June 14, 2019.

the existing information technology (IT) systems such as enterprise resource planning (ERP), manufacturing execution systems (MES), and hand-held personal digital assistant (PDA) devices. In most cases, the structural information is also stored alongside the production data and can be acquired from these IT systems. For instance, the process route can be determined from Bill of Material (BOM) and Bill of Process (BOP) data, which is often managed in ERP and MES. This structural information can be collected by converting data in IT systems to simulation data using appropriate data interfaces.

Unlike structural data, timing data used to estimate processing times is often difficult to obtain directly from IT systems and rarely produces accurate estimations of actual processing times especially in high-mix and low-volume production. For example, in many systems the time information is collected by workers manually indicating start and stop times of a process. This leaves many opportunities to introduce errors, such as forgetting to start/stop time collection or not appropriately accounting for interruptions in processing. Several researchers have addressed this issue. Hosoe et al. [4] proposed methods to estimate processing time from Point of Production (POP) data using regression techniques for semiconductor production systems. Meidan et al. [5] also reported methods for estimating processing time for semiconductor production systems. They enumerated the explanatory variables such as batch size and extracted the factors with a large influence on processing time using Bayes classifier. Karnok et al. [6] and Nagahara et al. [7] have proposed methods that estimate processing time from noisy (incomplete) POP data of real production systems. Throughout this paper, point of production (POP) data includes all data that is collected from the shop-floor during production, such start/completion times, and can be tied to a specific production process, resource, operator, etc.; the point where production occurred.

Operational control rules such as job sequencing rules and resource assignment rules determine dynamic behavior of production systems. Like processing time, operation control rules are rarely managed explicitly in IT systems. In most practical cases, operational control is executed according to human decisions and the rules underlying these decisions are tacit or hidden. Sprock et al. [8] have proposed the Operational Control Model as a standard and comprehensive representation of operational controls. They classified the operational controls into five categories: admission, sequencing, resource assignment, dynamic processing planning, and changing resource states. In production simulation, rules for each control function must be modeled appropriately. Therefore, in general, the rules are determined through by interviewing shop-floor managers.

In contrast to processing time, there is little research using data-driven approaches to modeling operational control rules. Bergman et al. [9, 10] proposed methods for identifying job sequencing rules from POP data using machine learning techniques. However, they evaluated the method through experiments using synthetic data generated from a simple simulation model. The applicability of this data-driven approach for more complex real-world scenarios has not been verified in this field.

Additionally, when operational control is conducted according to human decisions, uncertainty in the decisions will be a critical issue. Consider job sequencing as an example. When workers select the next job from waiting jobs, it is possible that the job with the highest priority may not be selected. This uncertainty in the job sequencing would make the rule identification inaccurate, and so it is important to consider this uncertainty in the rule identification process.

For the above issues, this research verifies the applicability of data-driven job sequencing identification methods through experiments using real production data. In addition, we propose a method that reduces the influence of the uncertainty in the job sequencing.

The remainder of this paper is organized as follows. In section 2, we describe the problem statement of job sequencing rule identification and review some related works. In section 3, we detail the proposed method. In section 4, we first evaluate the proposed methods on virtual production data and then show experimental results of applying the methods to real production data. Finally, in section 5, we discuss conclusions and future work.

2. Job Sequencing Rule Identification Problem

2.1. Problem Statement

Rule identification methods identify the rule mapping between system states and actions. For job sequencing rule identification, the job waiting to be processed corresponds to the state and selecting the next job to be processed corresponds to the action. The objective of the Job Sequencing Rule Identification Problem (JSRIP) is to derive a job ranking model that predicts the highest ranked job among the waiting jobs given the state/action data.

The systems of interest in this research are discrete production systems such as job-shop production systems in which job sequence is decided based on tacit rules. We assume that POP data is available for job sequencing rule identification.In addition, we assume that only one job is selected from waiting jobs and the selected job is processed immediately after selection. Additionally, a list of waiting jobs when each job selection event occurred can be derived from the POP data.

The job ranking model is trained using a training dataset composed with the event data. The accuracy of the trained model is evaluated using a test dataset composed with the other event data. We evaluate the accuracy of job ranking models using two metrics: Top-1 accuracy measures the rate at which the actual selected job (based on POP data) was predicted to be the best (Top 1) by the trained ranking model, and Top-5 accuracy measures the rate at which the rank of the actual selected job is predicted to be among the five best options.

2.2. Related Works on Ranking Method

JSRIP can be thought of as one form of a “Learning to Rank” problem. Learning to Rank problems are actively studied in the information retrieval field. The objective of Learning to Rank in that field is to extract documents with high relevance to a

Nagahara, Satoshi; Sprock, Timothy; Helu, Moneer.
”Toward data-driven production simulation modeling: dispatching rule identification by machine learning techniques.”
Paper presented at 52nd CIRP Conference on Manufacturing Systems (CMS 2019), Ljubljana, Slovenia. June 12, 2019 - June 14, 2019.

given query from a document set. A ranking model is derived from a dataset consisting of features of documents and queries. In the dataset, the degree of relevance of each document to the query is given as a label. Representative methods for Learning to Rank problems are divided into three approaches: pointwise, pairwise, and listwise [11].

Pointwise approaches look at a single document. A regressor or classifier is trained to predict the relevance of a particular document to a query. The input variables to the regressor/classifier are features of a single document-query pair, and the output is estimated relevance of the document to the query. This approach assumes that the relevance of each document is independent of the other documents in the document set. General regression and classification techniques can be used for this approach.

In pairwise approaches, a classifier is trained to estimate the magnitude relation of the relevance for a certain document-pair. Given a query and a pair of documents, the classifier is trained to minimize the difference between estimated magnitude relation of the documents and its ground truth. All the general classification techniques can be used for this approach, and the methods using Support Vector Machines, Boosting and Artificial Neural Network (ANN) as the classification technique are called as RankSVM [12], RankBoost [13] and RankNet [14], respectively.

Listwise approaches look at the entire list of documents. In thes approaches, a regressor is trained to estimate the score of a particular document for a query. The score is normalized by the scores of all documents in a query, and the loss is calculated based on the difference between the normalized score and the normalized relevance of each document. The representative method of this approach is ListNet that uses ANN as the classification technique [15]. In ListNet, the top one (Top-1) probability is calculated by normalizing the score/relevance by Soft-max function, and the loss is calculated from Cross Entropy Loss of the top one probability. Top-1 probability of a document is the probability that it is the best, or on top, from the list. Likewise, Top-*k* probability indicates it is ranked among the *k* best.

When formulating JSRIP as a Learning to Rank problem, the job selection event corresponds to the query and waiting jobs correspond to the documents. The characteristics of JSRIP compared to general Learning to Rank problems are as follows.

a) In JSRIP, the complete ranking of waiting jobs is unknown. We know only which job was selected. Therefore, the relevance is determined in two levels (e.g. 0 - 1 value). The relevance of the selected job is 1, and the relevance of the other jobs is 0.
b) In JSRIP, the relevance of each job is not independent of the other jobs. The job selection is decided based on the comparison of waiting jobs. From (b), it is obvious that pointwise approaches are not suitable for JSRIP.

The uncertainty in the job sequencing can be thought of as label noise in Learning to Rank problems. Niu et al. [16] investigated the influence of label noise in Learning to Rank problems. They concluded that fewer relevance levels and greater class imbalances increase the influence of label noise in the learning. From this perspective, JSRIP is a problem that is sensitive to label noise because of the problem characteristics (a). Therefore, it seems critical to consider how to reduce the influence of the uncertainty. Ding et al. [17] proposed a method for Learning to Rank problems with label noise. In their method, the reliability of each data sample is calculated using a generative model, then the loss function is weighted by the reliability. However, this method is not suitable for JSRIP because it is based on the pointwise approach.

As mentioned in the section 1, Bergmann et al. [9, 10] have proposed a job sequencing identification method. Their method is based on the pairwise approach and constructs a classifier that predicts the priority relationship between arbitrary two waiting jobs. They compared classification algorithms and data transformation techniques, and then verified the usefulness of their method through experiments using data generated from production simulation. We've not found any research that applied the listwise approach to JSRIP.

3. Proposed method

3.1. Feature Variables for Job Ranking Model

In most of practical production systems, jobs of similar product type tend to be processed consecutively to reduce setup times. While simple job sequencing rules such as first in first out (FIFO), earliest due date (EDD), shortest processing time (SPT), etc. are useful and well-accepted, the reduction of sequence-dependent setup operations is especially important in high-mix and low-volume production systems. To make these methods, such as Bergmann et al. [9, 10], applicable to more realistic scenarios, it is important to consider this context. Therefore, we include classification features capturing product type differences between waiting jobs and the previous (or in-process) job.

In general, each product type has categorical attributes such as product type name and numerical attributes such as product length. As a result, the features are determined as follows.

$$WT_i \coloneqq t^e - t_i^a \tag{1}$$

$$DD_i \coloneqq t_i^d - t^e \tag{2}$$

$$X_{i,k} \coloneqq \begin{cases} 1 & \text{if } x_{i,k} = x_k^p \\ 0 & \text{otherwise} \end{cases} \quad (k = 1, 2, \ldots, N_{cate}) \tag{3}$$

$$y_{i,k} \quad (k = 1, 2, \ldots, N_{nume}) \tag{4}$$

$$Y_{i,k} \coloneqq y_{i,k} - y_k^p \quad (k = 1, 2, \ldots, N_{nume}) \tag{5}$$

where WT_i is the waiting time of *i*-th waiting job in a certain event, and DD_i is the remaining time until due date of *i*-th job. And, t^e is the event occurrence date, t_i^a and t_i^d denote the arrival date and due date of *i*-th job respectively. Furthermore, $x_{i,k}$ is *k*-th categorical attribute value of product type of *i*-th job, and x_k^p is that of the job processed before the event. Likewise, $y_{i,k}$ is *k*-th numerical attribute value of product type of *i*-th job, and y_k^p is that of the job processed before the event. And, N_{cate} and N_{nume} denotes the number of categorical and numerical attributes, respectively, of each product type. The features shown in Eq. (3) and (5) will contribute to identify the rules including the setup reduction perspective.

Nagahara, Satoshi; Sprock, Timothy; Helu, Moneer.
"Toward data-driven production simulation modeling: dispatching rule identification by machine learning techniques."
Paper presented at 52nd CIRP Conference on Manufacturing Systems (CMS 2019), Ljubljana, Slovenia. June 12, 2019 - June 14, 2019.

3.2. Combination with Voting Filter

One idea to reduce the influence of the uncertainty is filtering unreliable samples from the training dataset. From this point of view, we propose a method that combines the ranking algorithms with Voting filter. Voting filter is one of the countermeasures for label noise in classification problems [18]. In this method, several weak classifiers are trained using different training datasets and/or different classification algorithms. Then, the reliability of each data sample is evaluated based on the predictions by those weak classifiers, and unreliable samples are filtered from the original dataset. There are two major voting algorithms to judge the reliability of each sample. One is *Majority* voting that judges the sample as reliable for which half or more weak classifiers correctly predicted the ground truth class. Second is *Consensus* voting that judges the sample as reliable for which all weak classifiers correctly predicted the ground truth class. Majority voting is generally considered better than Consensus voting [18].

In the listwise approach, the reliability judgement is conducted for each event data since the listwise approach looks at the entire list of waiting jobs. On the other hand, in the pairwise approach, the reliability judgement is conducted for each pairwise comparison data.

4. Experiment and Discussion

4.1. Experiment using Virtual Production Data

Since the ground truth rule is unknown in the real scenario, we start evaluating the proposed methods by conducting experiments using virtual production data. The virtual production data is created from a simulation model consisting of one machine. The machine processes 2,000 jobs sequentially. The attributes of each job such as product type, due date, arrival date and processing time are randomly set. At the machine, job selection is conducted based on the priority score shown in Eq. (6).

$$S_i = a \cdot WT_i + b \cdot \exp\left(-DD_i\right) + c \cdot \delta_i \tag{6}$$

where S_i, WT_i and DD_i denotes the priority score, waiting time and remaining time until due date of waiting job i respectively. If the product type of job i is the same as that of the job processed before, $\delta_i = 1$. Otherwise, $\delta_i = 0$. The first, second, and third terms corresponds to FIFO, EDD and setup-reduction rule respectively, and a, b and c are the weighting coefficients for each term. The value of these coefficients is randomly set so that each rule influences job selection in some extent.

In this experiment, two scenarios for job sequencing are considered. One scenario selects the job with the maximum score (maximum score selection scenario), and the second selects a job based on the probability shown in Eq. (7) (stochastic selection scenario).

$$P_i = \exp\left(S_i\right) \Big/ \sum_{k=1}^{M} \exp\left(S_k\right) \tag{7}$$

where P_i is the probability that job i is selected, and M denotes the number of waiting jobs. This probability is introduced to express the uncertainty in the job sequencing.

The flowchart of the experiments for the proposed method is shown in Fig. 1. In the experiments, we use RankNet and ListNet to compare the pairwise and listwise approaches. In these methods, ANN is used as a classifier/regressor. In addition, the methods with/without Voting Filter are also compared. In the training of ANN, the Cross Entropy Loss with the L2 regularization term is applied as the loss function [19]. The event data is divided into training, validation, and test datasets. To prevent data leak, the event data are arranged in order of event occurrence date, and then divided into the three datasets. The rate of events in each dataset is 40% for training, 20% for validation, and 40% for test. The validation dataset is used for tuning hyper-parameters of the L2 regularization term.

The experimental results using RankNet and ListNet are shown in Table 1. Cases 1 and 2 are the results from the maximum score selection and stochastic selection scenarios, respectively. Case 3 denotes the results when the training and validation dataset are generated from the stochastic selection scenario and the test dataset is generated from the maximum score selection scenario. In cases 1 and 2, there is no significant difference between RankNet and ListNet. However, in case 3, RankNet outperforms ListNet. This result means that in our case RankNet is more robust to the uncertainty.

One reason why RankNet is better may be related to the rate of samples that comply with the ground truth rule, i.e. correct samples. Consider an event where there are N jobs waiting and the job with the second largest score is selected. This event is regarded as an incorrect sample in the listwise approach. On the other hand, in the pairwise approach, N-2 samples among N-1 samples generated from the pairwise comparisons of the waiting jobs hold correctness. The actual rate of correct samples in the training dataset is 64 % and 15 % in RankNet and ListNet, respectively. These results suggest that the pairwise approach is suitable for JSRIP with the uncertainty compared to the listwise approach.

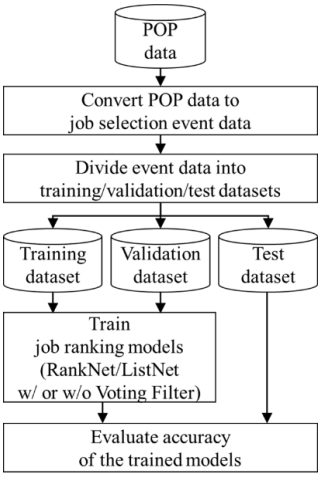


Fig. 1. Flowchart of the experiments for the proposed method

Nagahara, Satoshi; Sprock, Timothy; Helu, Moneer.
"Toward data-driven production simulation modeling: dispatching rule identification by machine learning techniques."
Paper presented at 52nd CIRP Conference on Manufacturing Systems (CMS 2019), Ljubljana, Slovenia. June 12, 2019 - June 14, 2019.

Table 1. Top-1 accuracy for virtual production data.

	Case	RankNet	ListNet
1	Maximum score selection	98.0 %	98.1 %
2	Stochastic selection	15.4 %	15.4 %
3	Training, Validation: Stochastic selection Test: Maximum score selection	88.9 %	71.5 %

Table 2. Result of reliable/unreliable judgement by Voting filter.

(i) RankNet

		Result of Voting Filter	
		Reliable	Unreliable
Correctness of data sample	Correct	(a) 62.6 %	(b) 1.6 % (2.4 %)
	Incorrect	(c) 1.5 % (4.1 %)	(d) 34.4 %

(ii) ListNet

		Result of Voting Filter	
		Reliable	Unreliable
Correctness of data sample	Correct	(a) 12.6 %	(b) 2.7 % (17.9 %)
	Incorrect	(c) 2.0 % (2.4 %)	(d) 82.7 %

Table 3. Top-1 accuracy of RankNet/ListNet for different sample sets.

	Case	RankNet	ListNet
1	All samples (a)(b)(c)(d)	88.9 %	71.5 %
2	Correct samples (a)(b)	97.5 %	91.8 %
3	Correct or reliable samples (a)(b)(c)	92.5 %	86.6 %
4	Correct and reliable samples (a)	92.5 %	81.9 %
5	Reliable samples (a)(c)	90.6 %	64.9 %

Next, the experimental results using RankNet/ListNet with Voting filter is shown in Table 2 and 3. Table 2 shows the results of reliable / unreliable judgement by Voting filter. The percentage values in this table denotes the rate of samples in the training dataset. For instance, the value in (a) denotes the rate of samples which are correct and judged as reliable by Voting filter. The values in the parentheses in (b) and (c) denote the false detection rate and overlooking rate respectively. The false detection rate is the rate of the samples judged as unreliable among all correct samples, and the overlooking rate is the rate of samples judged as reliable among all incorrect samples.

The reliable / unreliable judgment is made with a low false detection rate and overlooking rate in RankNet. On the other hand, the false detection rate is high in ListNet. This is because ListNet is a multi-class classification method, which is difficult to obtain the same prediction result among the weak classifiers, while RankNet is a 0-1 classification method. Table 3 shows the Top-1 accuracy of RankNet / ListNet for five cases in which the samples used for training are different. The case 5 corresponds to the result of the proposed method (RankNet / ListNet with Voting Filter). In the case 2 that only correct samples are used for the training, high accuracy is realized. And, the accuracy degrades due to increase of incorrect data (case 3), decrease of correct data (case 4), and both of them (case 5). This degradation is more pronounced in ListNet, and ListNet with Voting filter (case 5) is inferior to the original ListNet (case 1). On the other hand, RankNet with Voting filter shows the improvement compared to the original RankNet because the false detection rate and the overlooking rate are low as described above.

4.2. Experiment using Real Production Data

To evaluate the feasibility of the proposed method for realistic scenarios, we then conducted experiments using POP data collected from a real production plant. In this plant, over one hundred product types of industrial equipment are produced in mixed flow. The attributes of the product type are composed with eight categorical attributes such as product model name and five numerical attributes such as product length. From the POP data, we selected one process as a target and extracted about 2,000 job selection events in the process. In this process, jobs are processed by a machine, and a operator of the machine selects a next job from waiting jobs.

Table 4 shows the Top-1 and Top-5 accuracy of RankNet and ListNet. For the comparison, the accuracy when the rankings are predicted by general rules such as FIFO and EDD is also shown. As shown in the table, there is no significant difference in the prediction performance between RankNet and ListNet. In both methods, the Top-1 accuracy is not very high, but the Top-5 accuracy reaches about 90%. From this result, it can be considered that the trained ranking model can express the major tendency of the ground truth rule. One reason for the degraded Top-1 accuracy could be the uncertainty in the actual job sequencing, i.e. deviation from the ground truth rule.

Additionally, we investigate the feature importance to find the key features in the job sequencing rule. By using Random Forest as the classification method in the pairwise approach, the out-of-bag feature importance is calculated as shown in Fig. 2. The result shows that the key features are the waiting time (WT_A an WT_B in Fig.2) and the difference of 3rd and 4th categorical product type attributes between the waiting job and the job processed before (X_A3, X_B3, X_A4 and X_B4 in Fig.2). Here we have censored the actual attributes to protect the production data. This result indicates that the ground truth rule in this scenario includes the setup reduction perspective,

Table 4. Top-1 and Top-5 accuracy for real production data.

Ranking model	Top-1 accuracy	Top-5 accuracy
RankNet	61 %	90 %
ListNet	62 %	90 %
FIFO	5 %	36 %
EDD	15 %	49 %

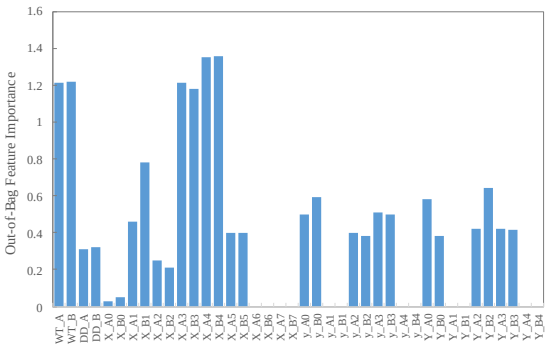


Fig. 2. Feature importance for real production data

Nagahara, Satoshi; Sprock, Timothy; Helu, Moneer.
"Toward data-driven production simulation modeling: dispatching rule identification by machine learning techniques."
Paper presented at 52nd CIRP Conference on Manufacturing Systems (CMS 2019), Ljubljana, Slovenia. June 12, 2019 - June 14, 2019.

and the features shown in Eq. (3), which describe the difference of categorical product type attributes between the waiting jobs and the job processed before, should be considered to identify the ground truth rule.

5. Summary and Conclusion

In this paper, we have proposed a job sequencing rule identification method to realize automated data-driven modeling of operational control rules for production simulation. At first, we organized the job sequencing rule identification problem (JSRIP) as a form of Learning to Rank problem. We clarified the difference between the two problems and investigated the applicability of Learning to Rank methods to JSRIP. Through the experiments using real production data, it was found that considering sequence-dependent setup operations is important for JSRIP in practical cases. By introducing the difference of product type attributes between the waiting job and the jobs processed before as features, the methods show good prediction accuracy for the real scenarios.

In addition, we compared the pairwise and listwise approaches for JSRIP with the uncertainty in the job sequencing. We found that the pairwise approach (RankNet) is more robust to uncertainty than the listwise approach (ListNet) and the pairwise approach can identify the ground truth rule accurately even in the existence of the uncertainty. Furthermore, we proposed a novel method utilizing Voting filter to reduce the influence of the uncertainty. Voting filter did not work well for the listwise approach. On the other hand, in the pairwise approach, the training samples were filtered with low false detection rate and low overlooking rate by Voting filter and the proposed method (RankNet with Voting filter) showed the improvement in the accuracy compared to the original RankNet.

The rules discovered by the proposed methodology are inputs into constructing production line simulation models. The rules are implemented by simulation queues that use the rules to sequence jobs to be processed. These simulations are useful for optimizing production planning and operations management decisions.

One future works is to evaluate using other ranking methods and filtering methods. It would be also interesting to investigate what kind of classification / regression algorithm is suitable for JSRIP with uncertainty. In addition, it is important to investigate how the prediction error of the job ranking model affects the accuracy of production simulation. Furthermore, rule identification methods for the other operation control rules such as resource assignment rules should be developed to automate the whole production simulation modeling process.

Disclaimer

Commercial equipment and materials might be identified to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the U.S. National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] Law, Averill M.. How to build valid and credible simulation models. Proceedings of the 2009 Winter Simulation Conference. 2009. pp. 24-33

[2] Fowler, J. W., Rose, O.. Grand challenges in modeling and simulation of complex manufacturing systems. Simulation. Vol. 80. Issue 9. pp.469-476.

[3] Barlas, P., Heavey, C.. Automation of input data to discrete event simulation for manufacturing: A review. Int. Journal of Modeling, Simulation and Scientific Computing. Vol.7. No.1. 1630001. 2016.

[4] Hosoe, H., Kanamori, N., Yoshida, K.. The methods of data collection and tool processing time estimation in lot processing. ISSM Paper. MC-P-075. 2007.

[5] Meidan, Y., Lerner, B., Rabinowitz, G., Hassoun, M.. Cycle-time key factor identification and prediction in semiconductor manufacturing using machine learning and data mining. IEEE transactions on semiconductor manufacturing. Vol. 24. No. 2. 2011. pp.237-248.

[6] Karnok, D., Monostori, L.. Extracting process time information from large-scale noisy manufacturing event logs. 44th CIRP Conference on Manufacturing Systems. 2011.

[7] Naghaara, S., Nonaka, Y.. Product-specific process time estimation from incomplete point of production data for mass customization. Procedia CIRP 67. 2018. pp.558-562.

[8] Sprock, T., McGinnis, L. F.. Simulation optimization in discrete event logistics systems: the challenge of operational control，Proceedings of the 2016 Winter Simulation Conference. 2016. pp.1170-1181.

[9] Bergmann, S., Feldkamp, N., Strassburger, S.. Approximation of dispatching rules for manufacturing simulation using data mining methods. Proceedings of the 2015 Winter Simulation Conference. 2015. pp.2329-2340.

[10] Bergmann, S., Feldkamp, N., Strassburger, S.. Emulation of control strategies through machine learning in manufacturing simulations. Journal of Simulation. 11. 2017. pp.38-50.

[11] Jagerma, R., Kiseleva, J., Rijke, M.. Modeling label ambiguity for neural list-wise learning to rank, arXiv:1707.07493v1. 2017.

[12] Herbrich, R., Graspel, T., Obermayer, K.. Support vector lerning for ordinal regression. Proceedings of ICANN. 1999. pp.97-102.

[13] Freund, Y., Iyer, R., Schapire, R. E., Singer, Y.. An efficient boosting algorithm for combining preferences. Proceedings of ICML. 1998. pp.170-178.

[14] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hunllender, G.. Learning to rank using gradient descent. Proceedings of ICML. 2005. pp.89-96

[15] Cao, Z., Qin, T., Liu, T., Tsai, M., Li, H.. Learning to rank: from pairwise approach to listwise approach. Microsoft TechReport. MSR-TR-2007-40. 2007.

[16] Niu, S., Lan, Y., Guo, J., Wan, S., Cheng, X.. Which noise affects algorithm robustness for learning to rank, Information Retrieval Journal. 2015. pp.215-245

[17] Ding, W., Geng, X., Zhang, X.. Learning to Rank from Noisy Data. ACM Transactions on Intelligent Systems and Technology. 2015

[18] Frenay, B., Verleysen, M.. Classification in the presence of label noise: a survey, IEEE transactions on neural networks and learning systems, Vol. 25, No. 5. 2014. pp.845-869

[19] Krogh, A., Hertz, J. A.. A simple weight decay can improve generalization. Advances in neural information processing systems. Vol. 4. 1992. pp. 950-957

Nagahara, Satoshi; Sprock, Timothy; Helu, Moneer.
"Toward data-driven production simulation modeling: dispatching rule identification by machine learning techniques."
Paper presented at 52nd CIRP Conference on Manufacturing Systems (CMS 2019), Ljubljana, Slovenia. June 12, 2019 - June 14, 2019.

Design Space Exploration for Wireless-Integrated Factory Automation Systems

Honglei Li*, Jing Geng*, Yongkang Liu†,
Mohamed Kashef†, Richard Candell†, Shuvra S. Bhattacharyya*
* University of Maryland, College Park, USA
† Intelligent Systems Division, National Institute of Standards and Technology, USA
Email: {honglei, jgeng}@umd.edu, {yongkang.liu, mohamed.hany, richard.candell}@nist.gov, ssb@umd.edu

***Abstract*—Recent years have brought significantly increased interest in integrating wireless communication capability within factory automation systems. Such integration motivates the study of interactions among the physical layout of factory workcells, wireless communication among workcells, and improving the overall factory system performance. In this paper, we develop methods for modeling and simulating these interactions, and implement these methods into the experimental study of complex design spaces for factory automation systems. The proposed methodology for modeling, simulation, and design space exploration can be used to gain insight into approaches for improving the configuration (e.g., physical layout or wireless protocol settings) of existing factory systems, and for understanding trade-offs in the design of new systems.**

I. Introduction

Modern factory automation systems are equipped with advanced wireless communications capability. Integration of such capability provides important potential advantages, such as lower cost to deploy and maintain networking capabilities within factories, and the ability to install sensors and monitoring functionality in parts of factories that are not possible to be efficiently instrumented using wired communications (e.g., see [1]).

Along with these potential advantages, integration of wireless communications introduces new challenges and novel constraints in the analysis and design of factory automation systems. A major source of these new challenges and constraints is the complex interaction among the factory layout and configuration. This interaction includes the placement of factory subsystems and their partitioning into nodes of the wireless network (network nodes); the performance of the wireless network that connects network nodes; and overall factory system performance. These factors lead to complex design spaces, which are composed of factory layouts, wireless communication networks, and interactions between them in system configuration and operations. We refer to these design spaces as *wireless-integrated factory system* (*WIFS*) design spaces.

In this paper, we develop new models and evaluation tools for understanding and experimenting with WIFS design spaces. Since evaluating these design spaces by physically constructing the different layout/networking combinations is in general infeasible, we present a new simulation-based design space exploration tool called *WISE* (*Wireless-Integrated factory System Evaluator*). WISE is designed for model-based simulation of factory automation subsystems that are equipped with wireless communications capability, and rapid simulation-based evaluation of alternative networked factory system designs.

Here, by *model-based*, we mean that the modeling techniques that underlie the tool are based on formal models of computation rather than on ad-hoc, tool-specific techniques that are difficult to precisely understand or to adapt to other modeling and simulation environments. Model-based design is a useful concept for many areas of cyber-physical systems and signal and information processing (e.g., see [2], [3]). The specific forms of model-based design emphasized in WISE are *dataflow modeling* for factory process-flows, and systematic interfacing of dataflow models with arbitrary network simulators that are based on discrete-event modeling.

The emphasis on dataflow is useful due to the utility of dataflow modeling across the areas of signal processing, control, and machine learning [3], which are all relevant to design and implementation of factory automation systems. This allows not only the high-level process-flow behavior of process networks to be modeled naturally and formally with WISE, but also lower level subsystems of the process-flows. Such a unified, model-based approach across levels of design hierarchy is useful for enhancing design modularity, analysis, and optimization.

Important features of WISE include capabilities for automatically generating (autogenerating) complex lower-level simulation models from compact representations at higher levels of abstraction. WISE also applies a new concept of *cyber-physical flow graphs* (*CPFGs*) as a graph-theoretic model for factory process-flows and other flow-oriented types of cyber-physical systems. We demonstrate WISE through extensive experiments that highlight its utility for exploring complex WIFS design spaces.

II. Background and Related Work

A significant body of the existing literature is relevant to modeling and simulation of factory automation systems that are equipped with wireless communication capabilities. Some of these works are based on novel applications of

Jeng, Jing; Li, Honglei; Liu, Yongkang; Hany, Mohamed; Candell, Richard; Bhattacharyya, Shuvra.
"Design Space Exploration for Wireless-Integrated Factory Automation Systems."
Paper presented at IEEE International Workshop on Factory Communication Systems (WFCS), Sundsvall, Sweden. May 27, 2019 - May 29, 2019.

existing simulation frameworks. For example, Liu et al. apply the OMNET++ simulation library to develop an integrated framework for factory process control simulation and wireless network simulation [4]. Marghescu et al. study the simulation of Zigbee-based wireless sensor networks using OPNET to evaluate and optimize the various network parameters [5]. Harding et al. develop a simulator that incorporates mathematical modeling and feedback control by developing an interface between MATLAB and OPNET [6].

Other works emphasize new models or simulation methods. For example, Vogel-Heuser et al. present approaches for modeling real-time requirements and properties of networked automation systems [7]. Schlick discusses advances, such as component-based automation and self-organizing production systems, in cyber-physical systems for factory automation [8]. Kurte et al. introduce a simulator for wireless sensor and actuator networks that allows simulation of heterogeneous systems through a novel interface abstraction for the operation of physical radio hardware [9]. Chaves et al. present a design environment for simulation and testing that is based on a service-oriented software architecture [10].

The novelty of the contribution in this paper centers on the development and application of WISE to explore complex WIFS design spaces. Compared to related work such as the works summarized above, distinguishing characteristics of WISE include its model-based architecture, which systematically integrates dataflow-based modeling of factory process-flows with discrete event modeling of wireless communication networks. WISE also provides autogeneration of low-level simulation code from high-level models, and cyber-physical flowgraph modeling, which further enhance the utility of the tool for WIFS design space exploration.

III. Design Flow

Fig. 1 illustrates the design flow associated with applying WISE for WIFS design space exploration.

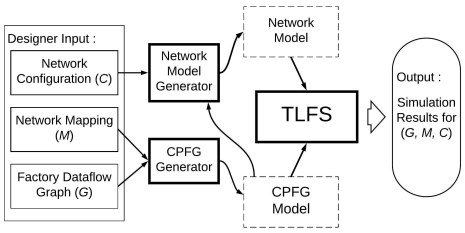


Fig. 1. An illustration of the new design flow involved in applying WISE for WIFS design space exploration.

As illustrated in Fig. 1, WISE builds upon a recently-introduced co-simulation tool called Tau Lide Factory Sim (TLFS) [11]. TLFS provides dataflow-based modeling of factory process-flows and systematic integration of the resulting process-flow models with arbitrary discrete event tools for network simulation. As illustrated in Fig. 1, WISE introduces and integrates with TLFS two new software tools, called the Network Model Generator and the CPFG Generator, and one new intermediate representation (graphical modeling data structure), called the CPFG Model. Additionally, the implementation of TLFS is extended in this work to support details of the CPFG Model. In Fig. 1, designer input, intermediate representations, and software tools are represented with thin-solid, dashed, and thick-solid borders, respectively.

A. Model-Based Architecture

The model-based architectures of WISE and TLFS emphasize dataflow-based modeling of factory process-flows and systematic interfacing between the process-flow models and arbitrary discrete-event simulators for communication architectures. Due to the abstract, model-based architectures of WISE and TLFS, the co-simulation and design space exploration techniques can be adapted readily to different dataflow-based design tools (for the process-flow modeling), and different communication network simulators.

A specific configuration of WISE involves two "plug-in" components for dataflow and communication network simulation. We refer to the two plug-ins as the *dataflow simulation plug-in* and *network simulation plug-in*, respectively. WISE systematically integrates the given pair of plug-ins into a model-based environment for exploring WIFS design spaces. In our experiments, which we report on in Section V, we utilize two specific dataflow and network simulation tools as plug-ins. These tools are, respectively, (1) the lightweight dataflow environment (LIDE) [12], and (2) the NS3 network simulation tool [13]. However, as described above, the model-based design of the WISE architecture enables retargeting the design space exploration techniques to other tools for dataflow and network simulation. This retargetability is useful because both of these areas for tool development — dataflow and network simulation tool development — are active areas for research and innovation.

B. Designer Input

The blocks in Fig. 1 labeled *Factory Dataflow Graph*, *Network Mapping*, and *Network Configuration* refer to simulation model input that is provided by the designer to represent the WIFS that is currently being studied.

The Factory Dataflow Graph models the factory process-flow between factory subsystems as a dataflow graph. The Factory Dataflow Graph is specified in a manner that is independent of the wireless network that is used for communication across distributed subsystems of the factory. Instead, the partitioning of Factory Dataflow Graph components into nodes of a wireless communication network, and the configuration of the network are specified separately. These specifications, represented by the blocks in Fig. 1 labeled Network Mapping and Network Configuration, are elaborated on in Section III-D. More details about Factory Dataflow Graph models are discussed in Section III-C.

The separation of concerns among the Factory Dataflow Graph, Network Mapping, and Network Configuration representations improves the efficiency and automation with which

Jeng, Jing; Li, Honglei; Liu, Yongkang; Hany, Mohamed; Candell, Richard; Bhattacharyya, Shuvra.
"Design Space Exploration for Wireless-Integrated Factory Automation Systems."
Paper presented at IEEE International Workshop on Factory Communication Systems (WFCS), Sundsvall, Sweden. May 27, 2019 - May 29, 2019.

the system designer can explore different ways of integrating wireless communication functionality into a given factory process-flow. In particular, the designer does not have to modify the Factory Dataflow Graph when the communication architecture changes; instead, only the relevant parts of the Network Mapping and Network Configuration specifications need to be changed. Then the detailed factory/communication co-simulation model is generated automatically. This separation of concerns and associated autogeneration capability is a major advance of WISE beyond TLFS.

C. Factory Dataflow Graphs

Formally, a Factory Dataflow Graph is a directed graph $G = (V, E)$, where the vertices (elements of V) represent factory subsystems such as machines, rails, parts generators, and machine/rail controllers. Directed edges (elements of E) in G represent the flow of information or physical entities (such as manufacturing parts) between factory subsystems. In the general terminology of dataflow graphs, the graph vertices are referred to as *actors*. Thus, actors in the Factory Dataflow Graph correspond to factory subsystems.

A dataflow graph executes by repeatedly executing actors that are ready (*enabled*) for execution, where the dataflow model provides a precise formulation for this form of readiness. As actors execute, they exchange packets of information (*tokens*) across the edges in the graph. These packets can have arbitrary data types associated with them, ranging from primitive types such as integers or floating point values to composite data types that correspond to user-defined objects (in an object-oriented programming sense). In Factory Dataflow Graphs, tokens may, for example, encapsulate information associated with the flow of physical parts, control messages, or instrumentation data.

Execution of a dataflow actor is decomposed into well-defined quanta of execution, called *firings*. Each firing is associated with characterizations of the amount of input data (number of tokens on the input edges) that is consumed by the firing, and the number of output tokens that is produced by the firing. These amounts of input and output data are referred to, respectively, as the consumption and production rates associated with the firing. An actor is said to be enabled for execution when there is a sufficient quantity of tokens buffered on its input edges, and a sufficient amount of empty buffer space available on its output edges to support the firing, as determined by the buffer sizes associated with the edges and the consumption and production rates of the firing.

For more background on the use of dataflow methods to model factory process-flows, we refer the reader to the detailed presentation of TLFS [11]. A notable difference, however, between the Factory Dataflow Graph of WISE and the dataflow graphs employed in TLFS is that Factory Dataflow Graphs do not incorporate any information about the communication architecture. These graphs are therefore simpler for the designer to work with. Furthermore, in conjunction with the separation of concerns described in Section III-B and the new automated model generation capabilities in WISE, Factory Dataflow Graphs are part of a more efficient approach for WIFS design space exploration. We elaborate on the automation capabilities further in Section IV, along with their utility in supporting design space exploration.

D. Network Mapping and Configuration

As shown in Fig. 1, the Network Mapping and Network Configuration are the two designer-provided inputs to specify the communication architecture that is to be integrated with the Factory Dataflow Graph for a given WIFS co-simulation (factory/network co-simulation). The Network Configuration input includes aspects related to factory layout.

Intuitively, the Network Mapping specifies how the given factory process-flow (as represented by the Factory Dataflow Graph) is distributed across different network nodes that communicate through wireless communication. The Network Mapping M for a Factory Dataflow Graph $G = (V, E)$ can therefore be represented as as a partitioning $M = N_1, N_2, \ldots, N_m$ ($m \geq 1$) of V — that is, the N_is are mutually disjoint subsets ($N_i \cap N_j = \emptyset$ for all $i \neq j$), and $N_1 \cup N_2 \cup \ldots \cup N_m = V$. To represent a fully centralized process-flow (with no wireless communication involved), one can simply set $m = 1$ so that the Network Mapping consists of just a single set $N_1 = V$. This type of mapping can be useful, for example, as a baseline to assess basic trade-offs associated with introducing wireless communication into the factory system.

The Network Configuration is another component of designer-provided input to WISE, as illustrated in Fig. 1. This input includes wireless communication parameter settings, such as the type of protocol and the propagation loss model. The desired Network Configuration settings are provided by the designer in a simple text file called `net_parameters.txt`. These parameter values are then converted to corresponding settings associated with the network simulation plug-in. To run a family of simulations with varying network parameters, the designer can easily edit the `net_parameters.txt` file or auto-generate a collection of files that can be iterated through for a set of simulation runs.

The parameters that can be specified in the `net_parameters.txt` file include the wireless communication protocol, propagation loss model, antenna transmitter gain, antenna receiver gain, noise figure for the noise signal, and others.

A Network Configuration specification for WISE also includes factory layout settings, which pertain to the spatial layout of factory subsystems, and can have significant impact on communication system performance. Factory layout settings in WISE network configurations are discussed in more detail in Section V.

E. Lower Level Models and Auto-generation

The input provided by the designer (user of WISE) is at a high-level of abstraction. This facilitates design space exploration because the models are easier to manipulate and reason about. However, to perform complete system simulation, the high level models must be translated into a lower-level form, which includes the simulation input to the network

Jeng, Jing; Li, Honglei; Liu, Yongkang; Hany, Mohamed; Candell, Richard; Bhattacharyya, Shuvra.
"Design Space Exploration for Wireless-Integrated Factory Automation Systems."
Paper presented at IEEE International Workshop on Factory Communication Systems (WFCS), Sundsvall, Sweden. May 27, 2019 - May 29, 2019.

simulation plug-in, and details of interfacing between the dataflow simulation plug-in and the network simulation plug-in. Such details are autogenerated in WISE by the blocks in Fig. 1 that are labeled Network Model Generator and CPFG Generator, respectively.

The output models that are generated by these two autogeneration subsystems are called the Network Model and CPFG Model, respectively. These two autogenerated models can be simulated together using WISE to achieve WIFS cosimulation between the given factory process-flow model and wireless networking capability that is integrated with the process-flow based on the given Network Mapping.

The structure and format of the generated network model are determined by the network simulation plug-in. As discussed previously, we presently employ NS3 as the network simulation plug-in. Thus, the Network Model Generator frees the designer from having to write NS3 code. The NS3 model is generated automatically from the designer's dataflow-based specification of the factory process-flow together with the Network Mapping and Network Configuration information.

The CPFG model includes special components, called *communication interface actors*, that model sending and communication of data between subsystems in a process-flow model. Communication interface actors model the exchange of data across a wireless communication network, and provide an abstract, modular interface between the dataflow simulation plug-in and the network simulation plug-in [11].

In Section IV, we discuss CPFG modeling concepts further, and provide an example of the CPFG model and parameterized network model that are generated from a given Factory Dataflow Graph and Network Mapping.

IV. Modeling and Autogeneration

In this section, we introduce details of the CPFG model, and its use as an intermediate representation in WISE. Second, we discuss communication link modeling for wireless channels in WISE. We also present a WIFS modeling example to illustrate the autogeneration of CPFG models and NS3 network models from the higher-level models provided as input to WISE.

A. Cyber Physical Flow Graph

The CPFG model is a specialized form of dataflow model that is useful for modeling and simulating WIFSs. In addition to its suitability for WIFSs, as we demonstrate in this paper, the CPFG model is applicable to a broad variety of modeling scenarios in cyber-physical systems. The CPFG model formulated here generalizes and formalizes an integrated, dataflow-based modeling approach for networked factory process-flows that was presented in preliminary form in [11].

Additionally, in this paper we introduce capabilities in WISE for autogenerating CPFG models from higher level representations. This is an important feature in streamlining the design process so that complex WIFS design spaces can be explored more efficiently, and more accurately.

A CPFG $G_{\text{cp}} = (V_{\text{cp}}, E_{\text{cp}})$ is a dataflow graph whose actors can be partitioned into three subsets V_p, V_c, V_i, which are called the physical, computational, and communication interface actors of G_{cp}, respectively. The computational actors correspond to actors in the usual sense of actors in signal processing oriented dataflow graphs (*dataflow process networks*) [14]. Such actors represent computational modules that represent discrete units of computation, called *firings*, as described in Section III-C.

Whereas an actor in a conventional signal processing oriented dataflow graph represents a computational module, a physical actor in a CPFG represents a physical subsystem or device, such as a factory machine or rail. A physical actor may encapsulate computational processing within it (e.g., processing that determines when to input a new part into a machine).

What distinguishes physical actors in the CPFG modeling approach is that any given physical actor must consume or produce *physical tokens* on at least one actor input or output, respectively. A physical token in turn models a discrete physical form of output (such as a generated or partially-processed part in a factory) rather than a packet of data, which is what a conventional dataflow token models. If a CPFG edge carries physical tokens, it is is referred to as a *physical edge*, otherwise, we call it a *cyber edge*.

As described in Section III-E, a communication interface actor (i.e., an element of V_i) models the sending or receiving of data across a communication network. In the CPFGs that we are concerned with in this work, the communication interface actors model wireless communication across distributed subsystems within a WIFS.

In WISE, communication interface actors provide a modular, model-based interface between the dataflow simulation plug-in and network simulation plug-in. For example, to retarget a CPFG to a different network simulator, one only has to change the implementations of the communication interface actor types. In WISE, we use only two types of communication actors, called *send interface actors* (*SIAs*) and *receive interface actors* (*RIAs*). Thus, only these two software components need to be retargeted to adapt a CPFG in WISE to work with a different network simulator.

As their names suggest, SIAs and RIAs model the sending and receiving of data, respectively, between dataflow actors across a communication network. For more background on SIAs and RIAs, we refer the reader to [11].

An example of CPFG modeling and associated use of SIAs and RIAs is presented in Section IV-C.

In summary, the CPFG model is distinguished by the partitioning of actors into physical, computational, and communication interface actors, and a dichotomy of edges as physical or cyber edges. A CPFG can apply general dataflow process networks [14] as the underlying dataflow model of computation or any specialized form of signal processing oriented dataflow that is compatible with the modeling requirements of communication interface actors. In this paper, we employ *core functional dataflow* (*CFDF*) [15] as the underlying dataflow model of computation. Background on CFDF and its utility in modeling factory process-flows is discussed in [11].

Jeng, Jing; Li, Honglei; Liu, Yongkang; Hany, Mohamed; Candell, Richard; Bhattacharyya, Shuvra.
"Design Space Exploration for Wireless-Integrated Factory Automation Systems."
Paper presented at IEEE International Workshop on Factory Communication Systems (WFCS), Sundsvall, Sweden. May 27, 2019 - May 29, 2019.

B. Communication Link Modeling

Fig. 2 illustrates different components of communication link modeling in WISE. Parameters associated with these components are configured by the designer as part of the Network Configuration block in Fig. 1, as described in Section III-D. Different antenna models are available for reception and transmission; the antenna is modeled as isotropic by default.

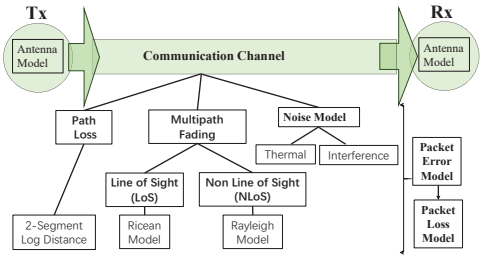


Fig. 2. Communication link modeling in WISE.

For the experiments reported on in this paper (Section V), signal noise is characterized as additive white Gaussian noise (AWGN). For the propagation loss model, a two-segment log distance model is applied. For multipath fading, Ricean and Raleigh models are used. For calculation of packet loss, the error rate is modeled based on a model presented by Miller [16], and subsequently validated by Pei and Henderson [17].

C. Autogeneration Example

In this section, we illustrate the models and autogeneration capabilities in WISE with a simple WIFS example.

Fig. 3 illustrates a Factory Dataflow Graph that is used to model a small-scale, pipeline-structured factory process-flow. The actor P represents a *parts generator*, which generates parts that are processed by the factory pipeline. The actors M_1 and M_2 model two machines that process parts, one by one, to add specific features to the parts. Parts are sent to and from each machine through rails, which are represented by the actors R_1 and R_2. The last stage in the pipeline is represented by the actor K. This actor, called the *parts sink*, represents a subsystem that collects and stores the parts after they are fully processed by the pipeline.

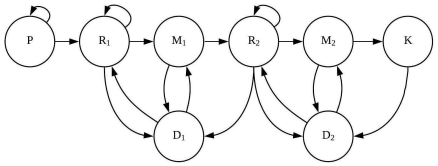


Fig. 3. An example of a Factory Dataflow Graph.

The actors D_1 and D_2 in Fig. 3 represent dual-rail, single machine (DRSM) controllers. A DRSM controller is a factory subsystem controller that is designed to interface with a single machine, a rail connected to the input of this machine, and a rail or parts sink that is connected to the machine output. Each DRSM controller sends commands to coordinate the flow of parts through the set of subsystems that it controls. For more details on the operation and modeling of DRSM controllers, we refer the reader to [11].

Fig. 4 illustrates the CPFG that is autogenerated by WISE for the Factory Dataflow Graph of Fig. 3 together with an example Network Mapping M. The mapping M involves seven distinct network nodes $N_1, N_2, \ldots, N_7$, and assigns the actors P, R_1, M_1, D_1, R_2, M_2, D_2, K, respectively to network nodes $N_1, N_1, N_2, N_6, N_3, N_4, N_7, N_5$. The solid edges in Fig. 4 carry physical tokens, while the dashed edges carry conventional dataflow tokens. In WISE, the determination of whether or not a given CPFG edge is a physical edge can be made automatically from the type of data that is associated with the Factory Dataflow Graph.

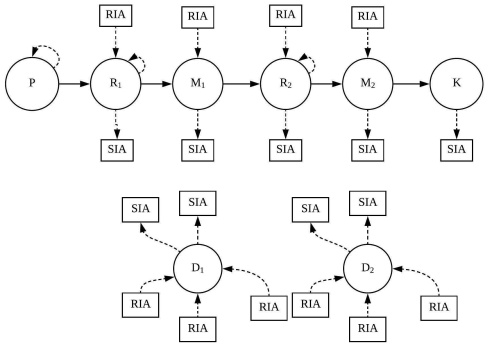


Fig. 4. Autogenerated CPFG.

The actors labeled SIA and RIA in Fig. 4 are communication interface actors that are automatically inserted by WISE in the process of autogenerating the CPFG. For each cyber edge whose source and sink actors are mapped to different network nodes, the communication associated with the edge is modeled with a separate (SIA, RIA) pair. For example, R_2 sends data to D_1, as shown by the edge (R_2, D_1) in Fig. 3, and these actors are mapped by M to distinct network nodes, N_3 and N_6, respectively. Accordingly an SIA S is connected to R_2 to model the sending of data to D_1 through a wireless channel, and a corresponding RIA is connected to D_1 to model the reception of data that is sent by S.

In WISE, all wireless communication is modeled in the autogenerated CPFGs through SIA-RIA pairs. Thus, all cyber edges in the CPFGs are associated with wired communication. In the current version of WISE, the latency of wired communication is assumed to be negligible compared to the latency of wireless communication and the execution time of machines. However, WIFS can readily be extended to incorporate latency models for wired communication — for example, by adding additional interfaces to the network simulation plug-in or by adding actors in the CPFG that model wired communication delays.

Fig. 5 illustrates the network model that is autogenerated

Jeng, Jing; Li, Honglei; Liu, Yongkang; Hany, Mohamed; Candell, Richard; Bhattacharyya, Shuvra.
"Design Space Exploration for Wireless-Integrated Factory Automation Systems."
Paper presented at IEEE International Workshop on Factory Communication Systems (WFCS), Sundsvall, Sweden. May 27, 2019 - May 29, 2019.

by WISE for the CPFG in Fig. 4. This graph shows the structure of the NS3 simulation model that is generated for co-simulation by TLFS with the generated CPFG. Each vertex N_i in Fig. 5 corresponds to a network node and each edge corresponds to a communication channel. The vertex Ap represents a single access point that is associated with the network nodes.

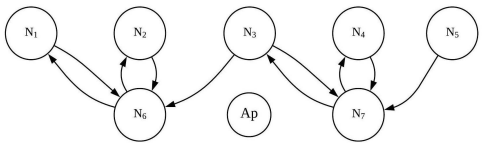


Fig. 5. The network model that is autogenerated by WISE for the example associated with Fig. 3 and Fig. 4.

Even for this simple, small-scale example, we see that the complexity of the CPFG together with the network model is significantly higher than that of the Factory Dataflow Graph, which is the designer's primary interface for working with WISE. This increase in complexity includes larger model sizes (more vertices and edges in the graph), as well as detailed software code that must be provided to correctly specify the lower-level models and ensure their consistency. The new models and autogeneration capabilities in WISE free the designer from the burden of managing this lower level design complexity.

V. Experiments

In this section, we demonstrate the utility of WISE through extensive experiments related to exploration of WIFS design spaces. We apply WISE in experiments with representative factory scenarios. Our experiments are performed using a desktop computer equipped with a 3.10 GHz Intel i5-2400 CPU, 4GB RAM, and the Ubuntu 16.04 LTS operating system.

A. Factory Layout Parameters

Presently, WISE assumes that a factory layout is in the form of one or more pipelines. Machines that belong to the same pipeline are arranged "horizontally", while different pipelines are arranged "vertically". Factory layout is therefore specified in terms of two distance-related parameters d_x and d_y, which respectively specify uniform (horizontal) spacing between successive subsystems (e.g., machines and rails) of a given pipeline, and uniform (vertical) spacing between successive pipelines in the vertical arrangement. Two additional layout-related parameters, N_p and N_m, specify the number of pipelines, and the number of factory machines within a given pipeline, respectively.

In most experiments in this section, we assume that each pipeline is assumed to have its own access point (AP), with a dedicated wireless channel assigned to each AP. It is assumed that if $N_p > 1$, then all of the dataflow occurs within the individual pipelines; that is, there is no communication across the pipelines. In Section V-G, we experiment with a set of scenarios in which all pipelines share a common access point.

The parameterized model of factory layouts supported in WISE represents a large class of factory systems with which capabilities of WISE can be demonstrated and experimented with. Also, the parameterized structure of the supported class of layouts is useful for demonstrating scalability-related factory performance trends. The extensible architecture of WISE makes it readily generalizable to support larger classes of factory layouts, such as layouts in which different pipelines have different numbers of machines, horizontal or vertical spacing between adjacent subsystems is non-uniform, or the overall layout structure does not necessarily involve horizontally-arranged pipelines. Such generalization is a useful direction for future work in WISE.

Fig. 6 shows an example of a factory layout of the form currently supported in WISE. In this example, $N_p = 2$, $N_m = 3$, and each pipeline has its own access point. Here, each $M_{i,j}$, $R_{i,j}$, and $D_{i,j}$ represents the jth machine, jth rail, and jth DRSM controller, respectively, for the ith pipeline. Each P_i, K_i, and A_i represents, respectively, the parts generator, parts sink, and access point for the ith pipeline.

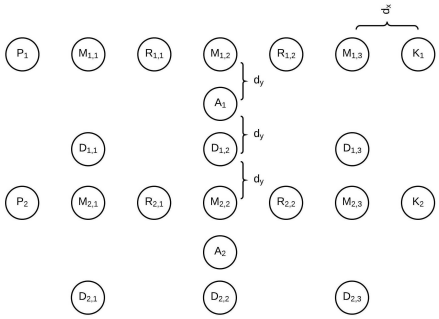


Fig. 6. Factory layout example.

B. Experiment Parameters

For each type of factory configuration simulated, we ran 50 WISE simulations independently and averaged the results. In each experiment, the simulation involved the production of 100 parts by each parts generator in each of the N_p pipelines, and the complete processing by the machines in each pipeline of the parts generated by the corresponding parts generator. The working time of each machine (the time required to process a given part) was determined randomly by the simulator using a designer-specified *mean working time* parameter μ. More specifically, the time for a given machine to process a given part was determined from a uniform distribution on $[0.9\mu, 1.1\mu]$. Each simulation terminated after the N_p parts sinks had each received 100 fully-processed parts.

The wireless communication protocol employed in all of the experiments reported on in this section is IEEE 802.11b. Since the protocol can be conveniently configured as part of the Network Configuration input to WISE, the experiments discussed here can be easily adapted to other protocols of interest.

Jeng, Jing; Li, Honglei; Liu, Yongkang; Hany, Mohamed; Candell, Richard; Bhattacharyya, Shuvra.
"Design Space Exploration for Wireless-Integrated Factory Automation Systems."
Paper presented at IEEE International Workshop on Factory Communication Systems (WFCS), Sundsvall, Sweden. May 27, 2019 - May 29, 2019.

WISE measures the communication delay associated with a packet P as $t_r(P) - t_s(P)$, where $t_s(P)$ is the time when P is sent by the corresponding SIA (see Section IV-A), and t_r is the time when P is received by the corresponding RIA. The average communication delay for a given simulation experiment is computed by averaging the difference $t_r(P) - t_s(P)$ over all communication packets.

C. Variation of Communication Delay with N_p

Fig. 7(a) shows how the average communication delay varies with the number of pipelines N_p. In this experiment, the Wi-Fi manager is configured to be the CARA (Collision-Aware Rate Adaptation) algorithm; $d_x = 10$ meters (m); $d_y = 10$ m; and $N_m = 3$.

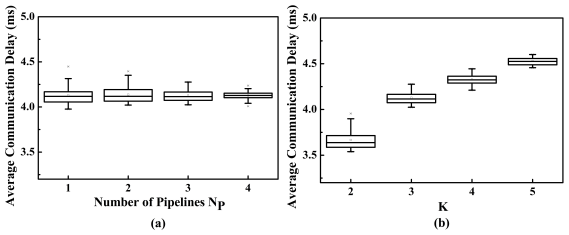


Fig. 7. (a) Variation in average communication delay with N_p. (b) Variation in average communication delay with $N_p = N_m = K$.

As shown in Fig. 7(a), the results for each N_p value are summarized in the form of a box plot. The endpoints of the vertical line segment for each plot extend from the minimum observed value to the maximum observed value. The three horizontal lines in each large box represent, from top to bottom, the 75th percentile, median, and 25th percentile of the corresponding set of 50 measurements. The small box inside each large box represents the mean value.

As shown in Fig. 7(a), the number of pipelines N_p has little influence on average communication delay for the class of factory systems considered in this experiment. This is because we allocate an independent access point for each pipeline and there is no communication between different pipelines.

D. Variation of Communication Delay with Both N_m and N_p

Fig. 7(b) shows results from an experiment where we have varied both the number of machines N_m and number of pipelines N_p. The variation is performed such that $N_m = N_p$. This allows us to visualize the effects of layout-complexity scaling in terms of a single parameter K, which is defined as the common value of N_p and N_m. The Wi-Fi manager algorithm is configured to be CARA as in Section V-C, and all other experiment parameters are as specified in Section V-B. The distance parameters are again configured as $d_x = 10$ m and $d_y = 10$ m.

As shown in Fig. 7(b), the average communication delay increases with larger K. This trend is largely due to two factors. First, the length of each pipeline increases with K, and correspondingly, the average distance from communication transceivers to the access point in each pipeline increases with K. Second, longer pipelines with more subsystems introduce more contention in the access points. The simulation results in this experiment provide specific insights on how communication delays vary and corresponding real-time performance issues are affected as a function of K, while other factory layout parameters are fixed.

E. Varying the Distance Parameters d_x and d_y

Fig. 8 presents a histogram of average communication delay, as determined by WISE simulation, with varying values of the distance parameters d_x and d_y. Each bar of the histogram is determined by averaging across 50 simulation runs. In this experiment, $N_p = 1$ and $N_m = 3$. The Wi-Fi manager algorithm is again configured to be CARA, and all other experiment parameters are as summarized in Section V-B.

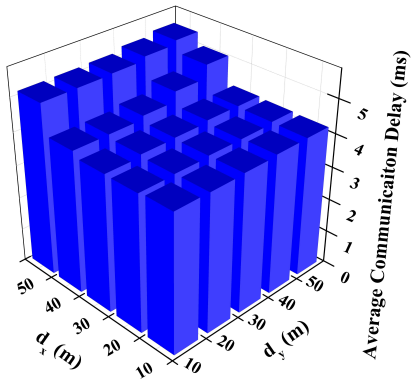


Fig. 8. Histogram of average communication delay with varying d_x, d_y.

This experiment shows a gradual trend toward increasing communication delay for $d_x, d_y \in \{10\text{ m}, 20\text{ m}, 30\text{ m}\}$, while for values of $d_x, d_y \in \{40\text{ m}, 50\text{ m}\}$, we see steeper rates of increase. We expect that this accelerated increase arises due to nonlinear effects such as the way in which the Wi-Fi manager downgrades the data rate when a significant frequency of communication failures is encountered.

F. Varying the Wi-Fi Manager Algorithm

Fig. 9 shows changes in the average communication delay with changes in the Wi-Fi manager algorithm and number of machines N_m. For these experiments, $N_p = 1$, and $d_x = d_y = 10$ m. All other parameters are set as summarized in Section V-B. The Wi-Fi manager algorithms investigated in this experiment are: Collision-Aware Rate Adaptation (CARA), Adaptive Auto Rate Fallback (AARF), collision detection for adaptive auto rate fallback (AARFCD), and Adaptive Multi Rate Retry (AMRR) [18].

G. Shared Access Point across Pipelines

In this section, we revisit the experimental setup of Section V-C with one change: we use a single, shared access point across all pipelines instead of a separate access point for each pipeline. Thus, the total number of access points in a given factory layout is reduced from N_p to 1.

Jeng, Jing; Li, Honglei; Liu, Yongkang; Hany, Mohamed; Candell, Richard; Bhattacharyya, Shuvra.
"Design Space Exploration for Wireless-Integrated Factory Automation Systems."
Paper presented at IEEE International Workshop on Factory Communication Systems (WFCS), Sundsvall, Sweden. May 27, 2019 - May 29, 2019.

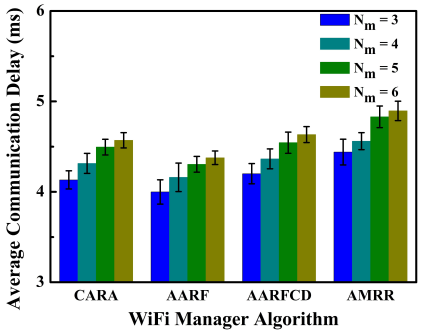


Fig. 9. Variation in average communication delay with the Wi-Fi manager algorithm and number of machines N_m.

As in Section V-C, the Wi-Fi manager algorithm is configured to be CARA; $d_x = d_y = 10$ m; and $N_m = 3$. All other experiment settings are as described in Section V-B.

Fig. 10 shows how the average communication delay varies with variation in the number of pipelines N_p under a single, shared access point configuration. We see in Fig. 10 a clear trend toward increasing average communication delay with increasing N_p. We anticipate that this is because with a single access point across all pipelines, increasing N_p results in more contention in the access point. Moreover, since d_x and d_y are fixed in this experiment, the average distance between communication transceivers and the access point increases with increasing N_p (see Fig. 6).

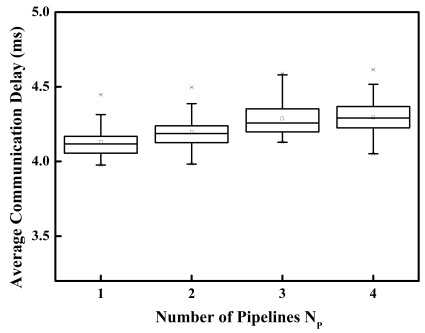


Fig. 10. Variation in average communication delay with N_p when a single, shared access point is used across all pipelines.

VI. Conclusions

In this paper, we have developed new models and computer-aided design tools that help in understanding and experimenting with complex, wireless-integrated factory system (WIFS) design spaces. Through extensive experiments, we have demonstrated the utility of our proposed new tools in exposing insights and performance trends involving multidimensional interactions among factory layout and communication system parameters. Useful directions for future work include developing optimization strategies, such as those based on randomized search (e.g., evolutionary algorithms or particle swarm optimization), for strategically iterating through families of simulations using our new WIFS-oriented models and tools.

Disclaimer

Certain commercial equipment, instruments, materials, software or systems are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] A. A. Kumar S., K. Ovsthus, and L. M. Kristensen, "An industrial perspective on wireless sensor networks — a survey of requirements, protocols, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1391–1412, 2014.

[2] E. A. Lee and S. A. Seshia, *Introduction to Embedded Systems, A Cyber-Physical Systems Approach*, 2011, `http://LeeSeshia.org`, ISBN 978-0-557-70857-4.

[3] S. S. Bhattacharyya, E. Deprettere, R. Leupers, and J. Takala, Eds., *Handbook of Signal Processing Systems*, 3rd ed. Springer, 2019.

[4] Y. Liu, R. Candell, K. Lee, and N. Moayeri, "A simulation framework for industrial wireless networks and process control systems," in *Proceedings of the IEEE World Conference on Factory Communication Systems*, 2016, pp. 1–11.

[5] C. Marghescu, M. Pantazica, A. Brodeala, and P. Svasta, "Simulation of a wireless sensor network using OPNET," in *Proceedings of the IEEE International Symposium for Design and Technology in Electronic Packaging*, 2011, pp. 249–252.

[6] C. Harding, A. Griffiths, and H. Yu, "An interface between MATLAB and OPNET to allow simulation of WNCS with MANETs," in *Proceedings of the International Conference on Networking, Sensing and Control*, 2007, pp. 711–716.

[7] B. Vogel-Heuser *et al.*, "Modeling of networked automation systems for simulation and model checking of time behavior," in *Proceedings of the International Multi-Conference on Systems, Signals & Devices*, 2012, pp. 1–5.

[8] J. Schlick, "Cyber-physical systems in factory automation — towards the 4th industrial revolution," in *Proceedings of the IEEE World Conference on Factory Communication Systems*, 2012.

[9] R. Kurte, Z. Salcic, and K. Wang, "A system level simulator for heterogeneous wireless sensor and actuator networks," in *IEEE International Conference on Emerging Technologies and Factory Automation*, 2018, pp. 776–783.

[10] A. Chaves *et al.*, "KhronoSim: A platform for complex systems simulation and testing," in *IEEE International Conference on Emerging Technologies and Factory Automation*, 2018, pp. 131–138.

[11] J. Geng *et al.*, "Model-based cosimulation for industrial wireless networks," in *Proceedings of the IEEE International Workshop on Factory Communication Systems*, 2018, pp. 1–10.

[12] S. Lin *et al.*, "The DSPCAD framework for modeling and synthesis of signal processing systems," in *Handbook of Hardware/Software Codesign*, S. Ha and J. Teich, Eds. Springer, 2017, pp. 1–35.

[13] *ns–3 Tutorial, Release ns–3.25*, ns–3 Project, 2016.

[14] E. A. Lee and T. M. Parks, "Dataflow process networks," *Proceedings of the IEEE*, pp. 773–799, May 1995.

[15] W. Plishker, N. Sane, M. Kiemb, K. Anand, and S. S. Bhattacharyya, "Functional DIF for rapid prototyping," in *Proceedings of the International Symposium on Rapid System Prototyping*, 2008, pp. 17–23.

[16] L. E. Miller, "Validation of 802.11a/UWB coexistence simulation," Tech. Rep., 2003.

[17] G. Pei and T. Henderson, "Validation of ns-3 802.11b PHY model," The Boeing Company, Tech. Rep., May 2009.

[18] *The ns–3 Wi-Fi Module Documentation*, ns–3 Project, 2016.

Jeng, Jing; Li, Honglei; Liu, Yongkang; Hany, Mohamed; Candell, Richard; Bhattacharyya, Shuvra.
"Design Space Exploration for Wireless-Integrated Factory Automation Systems."
Paper presented at IEEE International Workshop on Factory Communication Systems (WFCS), Sundsvall, Sweden. May 27, 2019 - May 29, 2019.

Analysis of Automatic through Autonomous – Ummanned Ground Vehicles (A-UGVs) Towards Performance Standards

Soocheol Yoon
Intelligent Systems Division
National Institute of Standards and Technology
Gaithersburg, MD, U.S.A.
soocheol.yoon@nist.gov

Roger Bostelman
Intelligent Systems Division
National Institute of Standards and Technology
Gaithersburg, MD, U.S.A.
roger.bostelman@nist.gov

***Abstract*— Automatic-through-Autonomous – Unmanned Ground Vehicles (A-UGVs), as termed by ASTM Standards Committee F45 for Driverless Industrial Vehicles, have much potential for use in manufacturing operations thanks to their versatility and flexibility. To utilize A-UGVs efficiently and effectively, it is needed to specify how the vehicle will be used, in what environment, and how best to control it. By understanding the detailed performance of the A-UGV and the facility environment, the vehicle can potentially operate with maximized productivity. In this paper, various parameters of the A-UGV are analyzed to measure navigation and obstacle avoidance performance. A-UGV aspects related to various facility environments are defined in structural form with organic relations. Performance test methods were developed and verified in a mock facility environment combining ramps and obstacles to measure navigation and obstacle avoidance performance.**

Keywords*— Automatic through Autonomous – Unmanned Ground Vehicles (A-UGVs), vehicle environment, vehicle performance, test methods

I. INTRODUCTION

Automatic through Autonomous – Unmanned Ground Vehicles (A-UGVs) is a relatively new term defined by ASTM Committee F45 standard F3200 as "automatic, automated or autonomous vehicle that operates while in contact with the ground without a human operator" [1][2]. A-UGVs can be applied in a range of facility types as next generation vehicles. They are deployed to perform transport and other operations for various applications, such as assembly, painting, maintenance, material handling, and military assistance [3-6]. The type of vehicle varies depending on the vehicle task. Vehicle types include: loading-carrying, typically called unit-load, forklift, or tugger which pulls carts. A relatively new type of tugger is called cart transporter where the cart straddles and is mechanically fixtured to the vehicle. "Dock" is defined in F3200 as the target location where the A-UGV interacts with another object – i.e., in this case an object can be a cart, conveyer, load, etc. Cart transporters push, pull, and rotate a cart docked to the vehicle with nearly the same motion as the mobile base, as opposed to a tugger A-UGV that pulls carts from a single pivot point (e.g., trailer hitch).

Each A-UGV type also has unique sensor configurations for use with their varied applications. A-UGVs may include horizontal and vertical laser sensors, sonar sensors, and other sensors as with the A-UGVs tested and described in this paper. As A-UGVs can have many potential uses, the specific purpose of the vehicle must be defined by the manufacturer or requested by the user so that the A-UGV specifications align with its intended use. For example, in a dynamic facility environment where moving obstacles may be present in the vehicle path, relatively long-distance clearances are required for front and side sensors, while relatively short clearances are required in narrow path driving. If both cases exist in the facility, the A-UGV sensor configuration must be able to handle the variations.

There has been much research to measure and enhance the performance of the A-UGV [7]. The performance criteria includes path planning, stability, robot coverage, navigation system, accuracy and repeatability, time duration, task completion, efficiency, dexterity, autonomy, and exploration of unknown environments [8-13].

To effectively operate the A-UGV in a factory, it is necessary to measure the performance of the A-UGV, describe the environment in which the A-UGV operates, and evaluate how the A-UGV responds to each environmental factor. For example, most A-UGV manufacturers only generally specify that their vehicle can or cannot respond to floor gaps, undulations, and grade changes in the manufacturing floor. However, these environmental conditions can critically affect the vehicle performance. The user may therefore have difficulty selecting an A-UGV considering the specific task and operating environment.

To ensure expected performance, the user must define and measure: 1) the A-UGV performance on the task, 2) the manufacturing environment factor where the A-UGV will operate, and 3) the performance of the A-UGV's response to manufacturing environmental factors. Accordingly, ASTM Committee F45 was developed to standardize how the various A-UGV aspects are defined and recorded during standardized tests [1]. This allows employing the same methods for testing vehicles so that manufacturers can describe their product's performance accurately and users and potential users can compare candidate A-UGVs to their task requirements. As the A-UGV term encompasses automatic guided vehicles (AGVs) and mobile robots, performance test methods within ASTM F45 are currently being developed to address generically automatic through autonomous industrial vehicles. Test methods tailored for specific vehicle types are left for future developments.

This paper defines criteria and describes the tests and results of A-UGV performance when they are used in mock facility environments. The criteria are defined as the environments in which the vehicle is to be driven and/or avoided. A description is provided for each criterion as to why each capability needs to be measured. Use cases were developed to demonstrate how the results can be applied in manufacturing processes or tested in even more complex environments that, for example, combine ramps and obstacles.

Yoon, Soocheol; Bostelman, Roger.
"Analysis of Automatic through Autonomous - Unmanned Ground Vehicles (A-UGVs) Towards Performance Standards."
Paper presented at 2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE 2019), Ottawa, ON, Canada. June 17, 2019 - June 18, 2019.

II. Driving performance

First, it must be considered that generally A-UGVs have different driving modes, including: speed-based; heading-based; purpose-based; autonomous driving; and/or automatic driving. In this paper it is assumed that the vehicle generally drives in autonomous driving mode with normal speed. Normal speed is the speed chosen by the vehicle controller up to a user-defined maximum that the vehicle will travel based on sensor and location inputs. Performance measurements of A-UGVs with specific mode configurations, such as maximum driving speed and acceleration are not considered in this study.

A. Narrow Path Driving and Curved Path Driving Performance

Paths that are slightly wider than the vehicle can define the test criteria for minimum passable area through which the A-UGV can navigate. Path width and navigation speed can be changed (dependent upon the facility and user requirements) according to vehicle configuration and capability. When moving from one path to a perpendicular path, as shown in Fig. 1, these criteria must be combined with curved path criteria.

Curved path driving performance describes how well the vehicle can follow an intended curved path and under what conditions – in this case, 90° to another path. Usually A-UGVs generate paths under the shortest path protocol resulting, in this case, in the curved path maintaining minimal clearance from the inside radius of the corner. In our tests, sometimes the vehicle, with a large turning radius, emergency-stops (or e-stops) when it reaches a specific proximity to the wall. When the A-UGV attempts to navigate the curve and detects the corner, the vehicle rotates with a small rotation radius, requiring more space and time to make the turn. The vehicle performance criteria can therefore be evaluated as measuring: maximum rotation radii, required inside clearance to the corner, and required outer area to complete the turn.

B. Grade Driving Performance

Grade, or ramp, is listed in ASTM F3218 as an environmental condition for A-UGVs, that affects mobility performance [14]. The primary performance criteria are whether the vehicle can: 1) navigate from a level surface and transition onto the ramp, 2) navigate on the ramp, and 3) transition onto a level surface from a ramp.

The secondary performance criteria include whether or not the A-UGV can identify a ramp when navigating on a level surface. When creating the vehicle navigation map, it is typically known where ramps are located. However, if the navigation system does not support ramps, ramps are typically

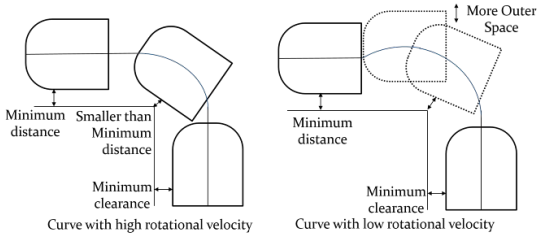


Fig. 1. Curved vehicle paths at high (left) and low (right) velocities.

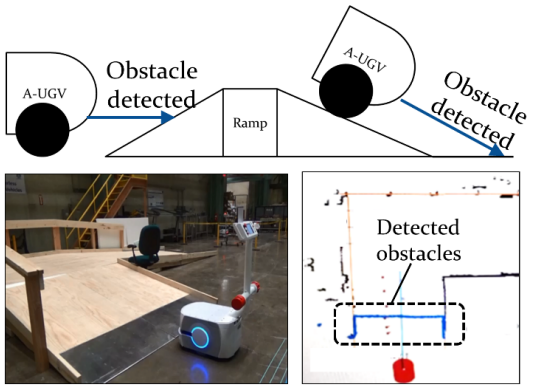


Fig 2. (top) Drawings of the ramp detected as an obstacle by the A-UGV. (bottom) Picture of the A-UGV in front of a ramp (left) and the ramp detected by the vehicle sensors (right).

detected as an obstacle by some sensors (e.g., line scanner range imagers) when the vehicle passes or enters the ramp. Similarly, when the vehicle travels down the ramp, the ground level surface is recognized as an obstacle by the front-facing laser sensor as shown in Fig. 2.

The common method to allow ramp-use by A-UGVs is to temporarily turn off low-mounted sensors which mainly detect the ramp as an obstacle when the A-UGV is on the level floor. Some other approaches can instead be applied to detect ramps both in software and hardware. For example, comparing the distance to the detected obstacle by various sensors can be used, and angle sensors can be used to measure vehicle tilt.

Therefore, grade driving performance can be described by: the ability to detect ramps, the ability to recognize being on a ramp; the required force or speed to enter a ramp; the minimum vehicle-to-ramp distance before transitioning onto a ramp; the transition capability from a ramp; the ability to drive on the ramp forward and backward; and the ability to rotate, drive, dock, and undock while on a ramp.

C. Driving Performance upon Gap and Threshold

Because not all factories are made up of just one uniform floor surface, there can be gaps and/or thresholds between surface areas. Separations between surfaces are gaps and adjacent floor height differences are thresholds. A crack in the factory floor is a typical example of a gap, irregular in shape. A concrete abutment with a grate inserted that is raised slightly (e.g., 1 cm) within a uniform floor surface is an example threshold.

Often, A-UGVs are expected to traverse gaps and thresholds. There are three possible results in this case: 1) the A-UGV successfully passes over it and detects or does not detect it (not detecting it could be considered a sensor failure), 2) the A-UGV fails to pass over it but does not detect it, or 3) the A-UGV fails to pass over it although recognizes it. The second case occurs when a vehicle wheel is stuck in the gap or at the threshold. In this case, the vehicle cannot move as planned although one or more driving wheels may keep moving. The vehicle controller inputs wheel encoder changes without vehicle body movement and as a result, vehicle position errors appear in the navigation system. Fig. 3 shows

Yoon, Soocheol; Bostelman, Roger.
"Analysis of Automatic through Autonomous - Unmanned Ground Vehicles (A-UGVs) Towards Performance Standards."
Paper presented at 2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE 2019), Ottawa, ON, Canada. June 17, 2019 - June 18, 2019.

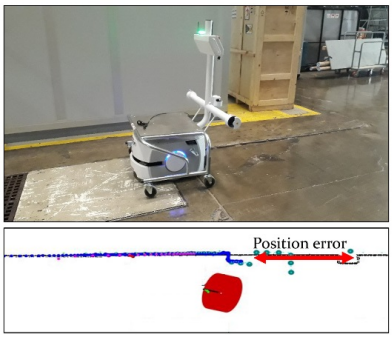


Fig. 3. A-UGV stuck on a raised surface resulting in vehicle position error.

an example of a position error from the A-UGV being stuck on a threshold.

Gap and threshold performance criteria can be described by: the ability to detect a gap or threshold, maximum length of a gap to pass over, maximum height of a threshold to pass over, minimum driving speed to pass over the maximum length of gap and height of a threshold, the ability to recognize position error caused by a gap or threshold, and the ability to recognize being stuck from a gap or threshold.

D. A-UGV Performance Caused by Floor Conditions

Factories can have different floor conditions. The most widely used floor materials are concrete and steel. However, as the types and purposes of factories vary, other floor materials, such as wood, tile, or carpet, may exist. Additionally, miscellaneous materials may be lying on the floor, such as plastic wrap, liquid, thin film, or tape. It may be intended that the A-UGV has the capability to drive over these materials. In which case, any of these floor conditions may cause the vehicle to perform unexpectedly.

A-UGV performance affected by floor conditions can be described by: the feasibility to drive on specific floor surfaces and the feasibility to drive over materials on floor surfaces.

III. Obstacle Avoidance performance

One of the most critical issues concerning A-UGVs is safety. To protect nearby humans, equipment, facilities, and the A-UGV itself, the A-UGV performance detecting obstacles and avoiding collisions is very important. However, if the vehicle is under strict obstacle detection and avoidance or is to stop when detecting an obstacle in the vehicle path, the A-UGV may fail to drive to the goal. To continue production, a trade-off between drive performance and safety may therefore occur. Of course, external approaches, such as facility-based obstacle sensors, can be effectively applied to enhance both performance and safety. Instead, this study aims to analyze the obstacle detection and avoidance issue from the perspective of the vehicle itself as discussed in this section.

A. General Obstacle

In ASTM F3200, obstacle, also referenced from ISO 8373, is defined as: static or moving object or feature, on ground, wall, or ceiling, that obstructs the intended movement [2][15]. A discussion that adjoins the obstacle definition also states: Ground obstacles include steps, holes, uneven terrain, and so forth. For this study, obstacles are considered to be every detectable object which is not registered in the A-UGVs navigation system.

Beyond preprogrammed automatic guided vehicles, A-UGVs with higher intelligence should be able to detect general obstacles and continue driving toward the goal while avoiding the detected obstacles. Usually there are three obstacle types as in the definition: 1) obstacles on the floor, such as workers, objects, products, pallets, or other A-UGVs; 2) obstacles attached to the wall, such as wall-mounted tools or open doors; and 3) obstacles suspended from the ceiling, such as an overhead crane. Fig. 4 shows typical examples of these.

For all three types, it is required that the A-UGV measure the distance to the obstacle in real time to decelerate and/or plan an appropriate path around the obstacle. The A-UGV should then determine whether to ignore the obstacle or not. Surely when the obstacle is far away from the path, the obstacle can be ignored. The obstacle can also be ignored when it is mounted on the wall or suspended from the ceiling if it is higher than the vehicle height. Industrial Truck Safety Development Foundation (ITSDF) B56.5 safety standard [16] states that test pieces testing the onboard sensors must be detected “within the contour area of the vehicle (including onboard payload, equipment, towed trailer and/or trailer payload).” Thus, the vehicle should be able to determine these conditions. As stated in the previous section, the ability for the A-UGV to distinguish obstacles and ramps from the environment is also important.

A-UGV performance criteria when detecting general obstacles can be described as: minimum and maximum detectable range from the vehicle, obstacle detection time, obstacle measurement uncertainty (i.e., obstacle size and location), and time spent detecting an obstacle. A-UGV performance criteria when avoiding a general obstacle can be described as: minimum required space to avoid an obstacle, the ability to determine avoidance feasibility, the ability to generate an alternate route, and the ability to stop before a collision.

B. Virtual Obstacle

There are areas where vehicles are prohibited, such as pedestrian walkways or human work areas. In these areas, although there is no physical object for sensors to detect, the vehicle is expected to avoid them. The virtual obstacle may therefore constrain the A-UGV path. A-UGV navigation software may be configured with forbidden lines or areas for the vehicle to avoid. As with general obstacles, the vehicle should keep a minimum clearance to the virtual obstacle. Also, when no path can be generated because of virtual obstacles, the vehicle should stop driving.

A-UGV performance criteria for nearby virtual obstacles can be described as: the ability to put virtual obstacles in the

Fig. 4. Typical examples of general obstacles: cone (left), open door (middle), crane hook (right).

Yoon, Soocheol; Bostelman, Roger.
”Analysis of Automatic through Autonomous - Unmanned Ground Vehicles (A-UGVs) Towards Performance Standards.”
Paper presented at 2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE 2019), Ottawa, ON, Canada. June 17, 2019 - June 18, 2019.

vehicle map and the ability to react correctly to virtual obstacles.

C. Negative Obstacle

Negative obstacles are areas where the floor is lower than the surface on which the A-UGV drives (e.g., a hole). Similar to ramps, ASTM F3218 considers negative obstacles as a ground surface condition, called a gap, where depth and length of the gap are recorded. However, ASTM F45 is considering whether to specifically define a negative obstacle or to add further condition information since gaps can have width that is larger than what may be considered a floor gap (e.g., manholes and loading docks). Since negative obstacles can cause severe damage to the A-UGV and without the obstacle being placed in the vehicle map as infrastructure to avoid, the A-UGV should detect and avoid negative obstacles that may be within its path.

The A-UGV needs to at least detect the length and width of the negative obstacle that is within the vehicle path to safely avoid it. Fig. 5 shows a negative obstacle caused by a grate removed from a manhole. Typically, the grate covers the hole and therefore it is not placed in the map. However, when the grate is removed the A-UGV should safely avoid it if the size of negative obstacle is correctly measured and placed in the vehicle map.

As discussed in section II-B, other types of negative obstacles (e.g., platform-to-ramp transitions) are generated by sensing changes to elevation in the flooring. When an A-UGV is on the level platform at the ramp top and ready to go down the ramp, the vehicle must detect that there are changes in floor level to vary vehicle velocity and/or apply downhill braking. Those floor height changes may be incorrectly sensed for both ramps and negative obstacles. However, the vehicle performance is dramatically affected where it should keep driving for the ramp case and should avoid the negative obstacle in that case.

A-UGV performance criteria when detecting a negative obstacle can be described by the ability to detect a negative obstacle and the minimum detectable length, width, and depth of a negative obstacle. A-UGV performance criteria when avoiding a negative obstacle can be the same as for the general obstacle. It may also be required to measure the vehicle performance that is capable of driving over the maximum length, width, and depth of the negative obstacle.

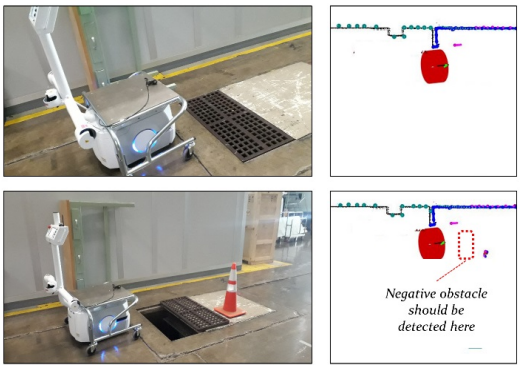


Fig. 5. A-UGV attempting to navigate across a covered (top) and uncovered (bottom) man-hole. The bottom image shows a negative obstacle not being detected by stock A-UGV sensors.

D. Atypical Obstacles

A-UGVs typically detect obstacles by light (e.g., laser detection and ranging (LADAR), vision) and/or acoustic sensors (e.g. sonar). LADAR and vision sensors are naturally affected by light. Similarly, acoustic sensors may be affected by loud noise. When bright light is directed or is flashed towards vehicle light sensors, the vehicle controller may interpret the light source as an obstacle as in Fig. 6. Common bright light examples in factories are uncovered windows that allow bright sunlight to pass through or high intensity work lights used by maintenance personnel. ASTM F3218 provides a brief description and method for recording several aspects associated with environment light conditions, including: ambient lighting type and source, direct highly-concentrated, directional lighting, ambient lighting source location, lighting intensity level, spectrum, and light exposure (i.e., continuous or transitional).

A-UGV performance criteria for robustness to light conditions could be described as maximum light intensity and light direction that the vehicle can withstand before it considers the light source as an obstacle.

E. Moving Obstacle

Detecting and avoiding moving obstacles requires more advanced technology as compared to stationary obstacles. Most common moving obstacles in a facility are humans and equipment (e.g., carts, forklifts).

Multiple A-UGVs can detect each other as moving obstacles and they can avoid one-another with appropriate moving-obstacle sensing and/or a central controller planning vehicle path as shown in Fig.7.

F. Overhanging Obstacle

Overhanging obstacles are general obstacles with obstacle parts detected or undetected causing the obstacle to appear to float in the path. Typical examples are folklift tines, pipe laying on and overhanging a cart, and stretched arms of worker. Because of the vehicle sensor detection capability and mounting location, the entire obstacle in the vehicle path may go undetected.

Overhanging obstacles may be undetected when: 1) objects are thin and long, 2) objects are mounted above or below A-UGV's horizontal sensors, and/or 3) objects are mounted between the A-UGV's vertical sensors. These conditions cause high collision risk with severe damage to the objects or A-UGV.

Fig. 8 shows overhanging obstacles at three different heights. Polystyrene obstacles were placed and an A-UGV was programmed to move straight forward toward each. Only the 2nd level obstacle was detected due to the vehicle sensor mounting height being in-line with the obstacle and the

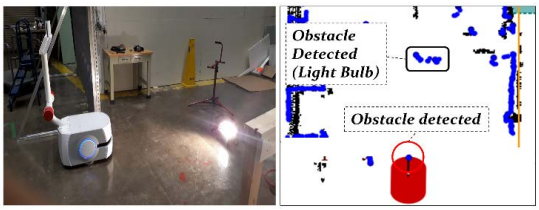


Fig. 6. Bright light detected as a (false) obstacle by the A-UGV obstacle detection sensors.

Yoon, Soocheol; Bostelman, Roger.
"Analysis of Automatic through Autonomous - Unmanned Ground Vehicles (A-UGVs) Towards Performance Standards."
Paper presented at 2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE 2019), Ottawa, ON, Canada. June 17, 2019 - June 18, 2019.

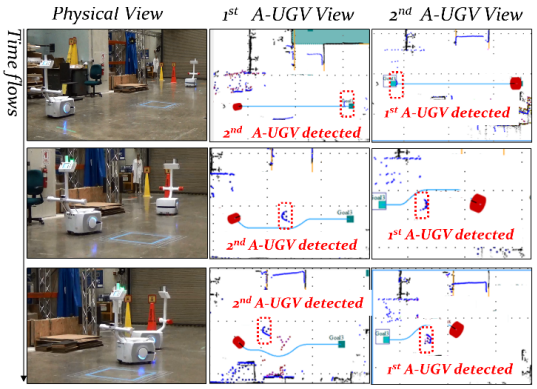


Fig. 7. Two vehicles approaching one another and recognized as obstacles by each other.

vehicle stopped before collision. The 1st and 3rd level obstacles were not detected causing A-UGV collisions with the obstacles.

Similar to general obstacles, A-UGV performance criteria with overhanging obstacles in the vehicle path can be described as the ability to detect obstacles that are within the facility – i.e., if there is a possibility of facility obstacles being within the A-UGV path, they must be detected and avoided. This includes the removal of sensor blind spots.

G. Small Obstacle

When two-dimensional (2D) A-UGV sensors are installed in only vertical or horizontal orientations, they may not detect obstacles that are shorter than, for example, a horizontal 2D LADAR sensor and narrower than the distance between right and left vertical sensors. Many obstacles fit this scenario (see Fig. 9), such as: worker tools and toolbox, cardboard packages, and tape rolls.

Sonar or other sensor types can be used to detect obstacles. Low-mounted sensors can detect obstacles just above the floor. Higher-mounted or adjustable-mounted sensors can detect obstacles in other locations.

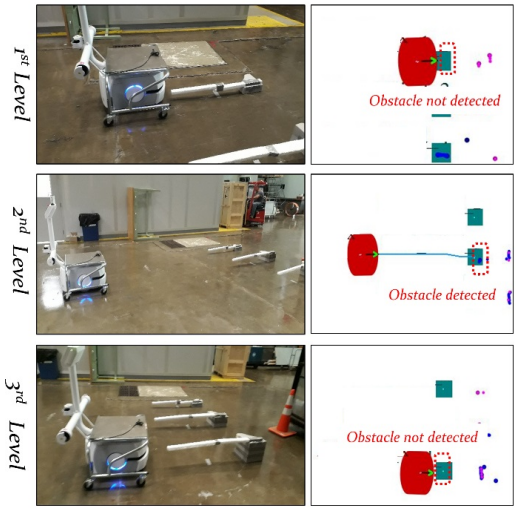


Fig. 8. Three overhanging obstacles with different heights being undetected (1st and 3rd level) and detected (2nd level) by the A-UGV sensors.

Fig. 9. A large, short box undetected by the A-UGV sensors.

As with overhanging obstacles, A-UGV performance criteria when detecting small obstacles can be described as the ability to detect obstacles that are within the facility.

IV. A-UGV Performance Case Study

In sections II and III, A-UGV performance considerations with respect to environmental effects and obstacles are discussed. There are many cases that combine these elements affecting A-UGVs. In some cases, there can be issues beyond those described previously that are caused by these complex conditions. In this section, a case study is presented where an A-UGV is programmed to avoid a general obstacle placed on a ramp and vehicle performance is measured.

A. Ramp with Obstacle

Given the information presented in section II-B, the addition of obstacles on the ramp can also affect vehicle performance, especially dependent upon their location on the ramp. Therefore, the effect of obstacle location will be analyzed.

The A-UGV was programmed to drive from the floor to the top platform of a 5° ramp measuring 2.4 m wide x 4.8 m long. An obstacle (cone) was then placed at 8 different ramp locations 1 m apart and from the top edge, namely Ob.1 to Ob.8, as shown in Fig. 10. The A-UGV was driven manually by the operator as a reference, and then 30 times in autonomous mode for the test. Manual mode includes direct human-machine interface (e.g., joystick) control and autonomous mode includes vehicle self-driving control. Throughout the test, 1) the A-UGV performance was measured and recorded, and 2) general issues were recorded when unexpected events occur.

The obstacle was placed in a position that can 1) interfere with A-UGV, and 2) allow enough space for the A-UGV to drive around. Each obstacle was therefore placed at 300 mm from the center of the ramp with the vehicle being aligned with the ramp center.

Through preliminary research, it was determined which A-UGV configurations affect driving performance. The A-UGV drives best on this ramp when 1) the low laser sensor obstacle detection is ignored, 2) vertical sensors are ignored, 3) local path fail distance is low, and 4) driving velocity is higher than 300 mm/s. These properties were applied and minimum driving velocity was set at 450 mm/s. Local path fail distance is the distance allowed to approach a sensed area blocked from navigation.

For each test, drive pass/fail, travel time, and travel distance were measured. A test was deemed as passing when the A-UGV successfully reached the goal. Otherwise, it was determined to be a failure. Overall A-UGV driving

Yoon, Soocheol; Bostelman, Roger.
"Analysis of Automatic through Autonomous - Unmanned Ground Vehicles (A-UGVs) Towards Performance Standards."
Paper presented at 2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE 2019), Ottawa, ON, Canada. June 17, 2019 - June 18, 2019.

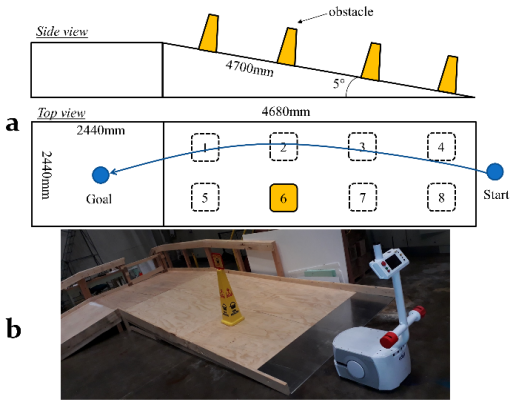


Fig. 10. (a) Drawings of eight different obstacles on a 5° ramp and (b) picture of an obstacle at location 6 on the ramp. The approximate A-UGV path avoiding the obstacle is shown from start to goal.

performance was evaluated, including: whether the vehicle could recognize the ramp as not being an obstacle, if it was able to transition to and from the ramp, and could drive on the ramp and avoid the general obstacle.

B. Results and Discussion

Table I shows the experimental results. In most cases, the A-UGV reached the goal. Failures occurred in only Ob. 5 and Ob. 6 obstacle locations - once for each. The A-UGV successfully drove up the ramp and approached the goal, although it kept spinning around the goal. These failures were self-errors of the A-UGV navigation system and can be treated as independent of the object avoidance performance.

The A-UGV therefore showed a success rate of 99.1% including the two failures. Compared with manual drive, travel time increased by 20.6 % and travel distance increased by 1.9 % on average. Ob. 8 and Ob. 4 showed the greatest increase in travel time, and Ob. 1 showed the least increase. Ob. 2 showed the largest increase in travel distance and Ob. 4 showed the least increase in travel distance.

Travel time was determined to be more critical than travel distance. There is no major difference in travel distance and the A-UGV drove in the allowed area for all cases.

Ob. 4 and Ob. 8 showed 40 % more travel time than the other obstacle locations. This was caused by the A-UGV's detection of the ramp and obstacle (see Fig. 11).

The ramp was detected as an obstacle crossing most of the available travel space. This was considered a general obstacle in the middle of the ramp but only occluded part of the ramp, allowing a path to be planned around the general obstacle and behind it in an S pattern, resulting in inefficient performance. It is possible to extend this issue of obstacles blocking other obstacles. For example, there can be a case that tricks the vehicle where it plans a path to a closed area.

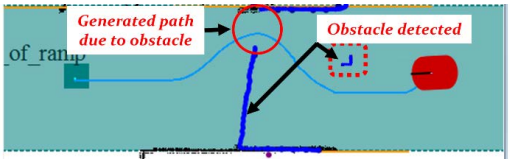


Fig. 11. Inefficient path generated due to obstacle (right) and ramp (left).

TABLE I. CASE STUDY EXPERIMENTAL RESULT

	Manual Drive (1 time)		**Autonomous Drive (30 times), mean [std]**	
	Time(s)	*Distance(mm)*	*Time(s)*	*Distance(mm)*
Free	16.3	6488	17.8 [0.8]	6491 [63]
Ob.1	18.2	6655	19.5 [1.1]	6734 [89]
Ob.2	16.4	6464	20.9 [1.5]	6802 [83]
Ob.3	16.9	6664	18.7 [0.5]	6697 [68]
Ob.4	17.5	6924	23.9 [2.1]	6872 [103]
Ob.5	16.8	6616	19.1 [0.9]	6757 [74]
Ob.6	16.8	6621	18.8 [0.7]	6722 [85]
Ob.7	16.6	6537	19.1 [1.6]	6658 [103]
Ob.8	17.1	6752	24.4 [5.3]	6999 [213]

The test case can be summarized as: 1) the vehicle was able to drive on the ramp with 99% success rate, 2) travel time was strongly influenced by the position of the obstacle while the travel distance does not change significantly, and 3) the travel time greatly increased when the obstacle was close to the ramp entrance.

A-UGV performance was measured through a series of experiments, and important issues were found for various situations. The A-UGV reacted differently depending on the circumstances (e.g., obstacle placement) and was verified that it can be measured quantifiably. The test was deemed successful for measuring the A-UGV performance using the testbed and test methods outlined.

The testbed can be expanded to a wider variety of situations. The configuration settings applied in this experiment were suitable for climbing the ramp, but there were several side effects. For example, it was difficult to find small and overhanging obstacles to place on the ramp because the lower and vertical sensors were not used.

V. CONCLUSION

This study analyzed an A-UGV's performance for typical facility applications particularly focusing on driving and obstacle avoidance. This study also considered the facility environment information for manufacturing system design and construction. Issues and requirements are defined to describe specifications of the A-UGV and facility environment. An A-UGV's performance was measured using a ramp with an obstacle in several locations on the ramp. Optimized configuration parameters were used while the test was performed. Results showed that for most trials the A-UGV drove well to the goal, although issues were uncovered when an obstacle was aligned with the ramp-detect location.

Standards for the A-UGV and its corresponding facility environment are under development by ASTM. For the negative obstacle and lighting issues discussed in this paper, specific hardware and software research activities are ongoing with help from the industrial and research communities. Future research should include a more detailed perspective of the facility environment. The A-UGV performance test methods should be standardized with the inclusion of irregular negative obstacles and dynamic negative obstacles (e.g., removable floor covers). Safety is an important issue and although there are many cases where safety and vehicle performance are conflicting, safety should be the primary consideration.

Yoon, Soocheol; Bostelman, Roger.
"Analysis of Automatic through Autonomous - Unmanned Ground Vehicles (A-UGVs) Towards Performance Standards."
Paper presented at 2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE 2019), Ottawa, ON, Canada. June 17, 2019 - June 18, 2019.

REFERENCES

[1] ASTM F45, Committee F45 on Driverless Automatic Guided Industrial Vehicles, https://www.astm.org/COMMITTEE/F45.htm

[2] ASTM, ASTM F3200 Standard Terminology for Driverless Automatic Guided Industrial Vehicles, April 1, 2018, DOI: 10.1520/F3200-18A

[3] Hamner B, Koterba S, Shi J, Simmons R, Singh S (2010) An autonomous mobile manipulator for assembly tasks. *Auton Robot* 28(1):131-149. http://dx.doi.org/10.1007/s10514-009-9142-y

[4] Shunan Ren, ying Xie, Xiangdong Yang, Jing Xu, Guolei Wang, Ken Chen, A Method for Optimizing the Base Position of Mobile Painting Manipulator, IEEE Transaction on automation science and engineering, Vol 14, No1, Jan 2017

[5] Pin FG, Culioli JC (1992) Optimal Positioning of Combined Mobile Platform-Manipulator Systems for Material Handling Tasks. *J Intell Robot Syst* 6(2-3):165-182. http://dx.doi.org/10.1007/BF00248014

[6] Najjaran H, Goldenberg A (2007) Real-time motion planning of an autonomous mobile manipulator using a fuzzy adaptive Kalman filter. *Robot Auton Syst* 55(2):96-106. http://dx.doi.org/10.1002/j.robot.2006.07.002

[7] Bostelman, R., Hong, T., & Marvel, J. (2016). Survey of research for performance measurement of mobile manipulators. Journal of Research of the National Institute of Standards and Technology, 121, 342-366

[8] Tang CP, Miller PT, Krovi VN, Ryu JC, Agrawal SK (2011) Differential-Flatness-Based Planning and Control of a Wheeled Mobile Manipulator - Theory and Experiment. *IEEE/ASME Transactions on Mechatronics* 16(4):768-773.

[9] Wong SC, Middleton L, MacDonald BA, Auckland N (2002) Performance metrics for robot coverage tasks. In *Proceedings of Australasian Conference on Robotics and Automation*, vol 27, pp 29.

[10] Papadopoulos EG, Rey DA (1996) A new measure of tipover stability margin for mobile manipulators. In *Proceedings of the 1996 IEEE International Conference on*, vol 4, pp 31S11-3116. http://dx.doi.org/10.1109/ROBOT.1996.509185

[11] Papadopoulos E, Rey DA (2000) The force-angle measure of tipover stability margin for mobile manipulators. *Veh Syst Dyn* 33(1):29-48. http://dx.doi.org/10.1076/0042-3114(200001)33:1;1-5;Ft029

[12] Sprunk C, Röwekämper J, Parent G, Spinello L, Tipaldi GD, Burgard W, Jalobeanu M (2014) An Experimental Protocol for Benchmarking Robotic Indoor Navigation. In *Proceedings of the International Symposium on Experimental Robotics (ISER)*.

[13] Thibodeau BJ, Deegan P, Grupen R (2006) Static analysis of contact forces with a mobile manipulator. In *Proceedings of the 2006 IEEE International Conference on Robotics and Automation*, pp 4007-4012. http://dx.doi.org/10.1109/ROBOT.2006.1642317

[14] ASTM, ASTM F3218 Standard Practice for Recording Environmental Effects for utilization with A-UGV Test Methods, July 1, 2017

[15] International Organization for Standardization (ISO), ISO 8373 Robots and robotic devices – Vocabulary, March 1, 2012

[16] ITSDF B56.5 Safety Standard for Driverless, Automatic Guided Industrial Vehicles and Automated Functions of Manned Industrial Vehicles, www.itsdf.org, accessed February 2, 2019.

Yoon, Soocheol; Bostelman, Roger.
"Analysis of Automatic through Autonomous - Unmanned Ground Vehicles (A-UGVs) Towards Performance Standards."
Paper presented at 2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE 2019), Ottawa, ON, Canada. June 17, 2019 - June 18, 2019.

Proceedings of the ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019

June 10-14, 2019, Erie, PA, USA

MSEC2019-2921

WHERE DO WE START? GUIDANCE FOR TECHNOLOGY IMPLEMENTATION IN MAINTENANCE MANAGEMENT FOR MANUFACTURING

Michael P. Brundage*
Thurston Sexton
National Institute of Standards and Technology
Gaithersburg, MD 20814, USA

Melinda Hodkiewicz
University of Western Australia
Crawley WA 6009, AUS

KC Morris
National Institute of Standards and Technology
Gaithersburg, MD 20814, USA

Jorge Arinez
GM Research and Development Center
Warren, MI 48090, USA

Farhad Ameri
Texas State University
San Marcos, TX 78666, USA

Jun Ni
University of Michigan
Ann Arbor, MI 48109, USA

Guoxian Xiao
GM Research and Development Center
Warren, MI 48090, USA

ABSTRACT

Recent efforts in Smart Manufacturing (SM) have proven quite effective at elucidating system behavior using sensing systems, communications and computational platforms, along with statistical methods to collect and analyze real-time performance data. However, how do you effectively select where and when to implement these technology solutions within manufacturing operations? Furthermore, how do you account for the human-driven activities in manufacturing when inserting new technologies? Due to a reliance on human problem solving skills, today's maintenance operations are largely manual processes without wide-spread automation. The current state-of-the-art maintenance management systems and out-of-the-box solutions do not directly provide necessary synergy between human and technology, and many paradigms ultimately keep the human and digital knowledge systems separate. Decision makers are using one or the other on a case-by-case basis, causing both human and machine to cannibalize each other's function, leaving both disadvantaged despite ultimately having common goals.

A new paradigm can be achieved through a hybridized systems approach — where human intelligence is effectively augmented with sensing technology and decision support tools, including analytics, diagnostics, or prognostic tools. While these tools promise more efficient, cost-effective maintenance decisions, and improved system productivity, their use is hindered when it is unclear what core organizational or cultural problems they are being implemented to solve. To explicitly frame our discussion about implementation of new technologies in maintenance management around these problems, we adopt well estab-

*Corresponding Author: mpb1@nist.gov

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

lished error mitigation frameworks from human factors experts — who have promoted human-systems integration for decades — to maintenance in manufacturing. Our resulting tiered mitigation strategy guides where and how to insert SM technologies into a human-dominated maintenance management process.

1 INTRODUCTION

The era of big data and Internet of Things (IoT) in manufacturing – with low cost sensors and cloud-based solutions – has left many manufacturers with a plethora of data in many different forms. With the recent buzz around more accessible, easy to use Artificial Intelligence (AI) solutions, some manufacturers are asking themselves "can we just throw our data in an AI?" Other manufacturers might ask "how can we get smart with new technologies?" However, AI and other Smart Manufacturing (SM) technologies are not one-size-fits-all solutions for all data types or problems, especially when there are many human-centered aspects in the workflow. Most AI and digital solutions do not work out-of-the-box and cannot directly replace personnel in many situations. Within manufacturing, maintenance is inherently one of the most human-centric processes, but is uniquely suited for an approach designed to intertwine human and digital capabilities. A new paradigm is needed that involves the human, AI and other advanced technologies working collaboratively and efficiently within the maintenance workflow. Achieving this paradigm requires an understanding of how and why this fails to happen in current maintenance practice. This paper dissects the maintenance workflow into the tasks that are performed by personnel, so that commonly occurring errors can be analyzed in a unifying error framework. Using this framework enables a tiered approach to technology implementation, guidance which is useful when manufacturers do not necessarily know where to start.

The rest of the paper is structured as follows: the remainder of Section 1 discusses maintenance in manufacturing, including the maintenance management workflow, maintenance strategies, and issues that occur in practice; Section 2 presents well established human factors research and how it will be applied to maintenance in manufacturing; in the subsequent sections, the maintenance workflow is broken down into three high level tasks: 1. Preparing for Maintenance (Section 3), 2. Performing Maintenance (Section 4), and 3. Discovering Maintenance Needs (Section 5). Within each of these sections, the applicable research and technologies are discussed, with high level tasks being further decomposed into sub-tasks to determine the types of errors that can occur. At the end of each subsection, the mitigations for these different example errors are discussed. Section 6 summarizes the steps from Sections 3-5 to generalize the error mitigation so manufacturers can follow similar steps. Lastly, Section 7 presents conclusions and future work opportunities.

1.1 Maintenance in Manufacturing

Maintenance is a collection of "actions intended to retain an item in, or restore it to, a state in which it can perform a required function" [1]. It is estimated that in 2016, US manufacturers spent $ 50 billion on maintenance and repair, which is between 15 % and 70 % of the cost of goods produced [2]. This estimate includes outsourcing of maintenance and repair, but does not include expenditures on labor and materials or the value of lost productivity due to unscheduled downtime. Estimates suggest that employing smart technologies can reduce maintenance cost from 15 % to 98 % with a high return on investment (ROI) [2]. Within the aerospace industry, examples of specific savings include an estimated return on investment of 3.5:1 for moving from reactive to predictive maintenance on electronic display systems [3] and a 56 % savings in costs from switching from reactive to predictive maintenance for train car wheels [4, 5].

The practice and delivery of maintenance has evolved over the last fifty years. During the late 1960's Nolan and Heap's [6] investigation of failures in the airline industry led to the development of reliability-centred maintenance (RCM), a process still widely used today. Building on RCM, a well defined theoretical and practical structure for maintenance management now exists. This is documented in standards [7], textbooks [8, 9, 10] and by professional societies [11, 12].

In the 1970s Japanese manufacturers introduced the concept of Total Productive Maintenance (TPM) [13]. The elements of TPM are 1) a focus on maximizing equipment effectiveness, 2) establishing a system of preventive maintenance for the equipment's entire life, and 3) the participation of all employees in maintenance through a team effort with the operator being responsible for the ultimate care of his/her equipment [14]. TPM is widely adopted in mature manufacturing organizations with well documented benefits [15]. While RCM and TPM are not competing frameworks, they have different goals: RCM determines an appropriate maintenance strategy while TPM is concerned with managing how maintenance is executed.

In the late 1990s Lean maintenance became popular, which built on the concepts of TPM and RCM and promised a transformation in manufacturing management through standardized workflows, value stream mapping, just-in-time (JIT) and Kanban "Pull" systems, Jidoka (Automation with a human touch), Poka Yoke (Mistake proofing), and the use of the plan-do-check-act process [16]. Despite the promised benefits of lean maintenance mentioned earlier, a literature review by Mostafa et al. found that research on applying lean principles into maintenance had not provided convincing evidence of success [17].

1.2 Maintenance Strategies

Two artifacts of the maintenance management system are particularly noteworthy. First is the *maintenance work order* (MWO). This concept refers to the archival record of the main-

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

tenance event from its inception to its completion and is shared along the way throughout the workflow. All maintenance work should be associated with a work order. The second concept is the *Computerized Maintenance Management System (CMMS)*. This system supports maintenance management with a record of the maintenance work orders and through access to documentation of the assets, resources, and other relevant information. In the workflow we present here, the CMMS is a hypothetical system and any actual implementation will vary. A MWO is generated, tracked, and eventually archived in the CMMS. The CMMS generates reports documenting the tasks that are due. The maintenance strategies are dependent on when MWOs are acted on and how they are planned through the CMMS.

Preventive maintenance is defined as the "actions performed on a *time-* or *machine-run-based* schedule (sometimes referred to as interval based) that detect, preclude, or mitigate degradation of a component or system with the aim of sustaining or extending its useful life through controlling degradation to an acceptable level" [12]. Preventive tasks and intervals are often found in manuals from original equipment manufacturers and are usually a requirement as part of warranty. Over time many asset management organizations develop their own preventive maintenance tasks and intervals as they gain knowledge about their assets and systems.

Condition-based maintenance is defined by Society of Maintenance and Reliability Professionals (SMRP) [12] as "an equipment maintenance strategy based on measuring the condition of equipment against known standards in order to assess whether it will fail during some future period and taking appropriate action to avoid the consequences of that failure. The condition of the equipment can be measured using *condition monitoring, statistical process control, equipment performance or through the use of human senses*." Maintenance personnel have been using inspections, process variables, vibration analysis, thermography, oil analysis, ultrasonic analysis, and other techniques for over 30 years. Predictive maintenance is defined in this paper as involving physical, statistical, or machine learning models that combine historical reliability and/or performance data with current condition assessment to generate a probability of failure and/or failure event prediction interval. These machine learning models are used to support condition-based maintenance programs and to inform interval selection for preventive maintenance tasks.

Despite best efforts at proactive maintenance, the stochastic nature of asset degradation means that failures do occur and reactive maintenance is necessary. These failures result in corrective work, which as will be seen in detail below disrupts the maintenance management process. Depending on the consequence of the failure, corrective work may need to be executed immediately (unstructured work). Otherwise, work can be passed into the planning process (structured work). In the manufacturing domain, corrective work is often referred to as unstructured reactive work [2].

1.3 Maintenance Management Workflow

A major factor for the efficiency of maintenance management is whether the work is structured or unstructured. To describe the preferred maintenance structure, reliability engineers have broken down the maintenance workflow into six major steps: 1) Analyze, 2) Select & Prioritize, 3) Plan, 4) Schedule, 5) Execute, and 6) Complete.

Analyze The Analyze activity relies on the data documented in the work order. Planners, maintenance and reliability engineers use this data to inform their respective tasks. These include reviewing inspection and as-found condition reports to determine whether asset deterioration meets expectation and when the asset has deteriorated past that expected threshold reviewing existing strategy or interval settings for inspection and maintenance, updating data for reliability and risk calculations, and updating optimization models. Analysis is involved in many of the maintenance management processes.

Select & Prioritize Maintenance work can be identified by many agents, such as operators, maintainers, engineers, and data analysts, by events (e.g., safety incidents), as well as from strategies stored in the CMMS, and in asset management plans, which include recommended routine maintenance schedules. There is always more work to do than can be done in any one planning and scheduling period, and hence work needs to be prioritized. Ideally there should be a risk-based process to prioritize work for each planning cycle. New work notifications arrive each cycle and are reviewed alongside work orders already on the backlog and scheduled preventive maintenance work orders due in the next maintenance work cycle. From these work orders a list of tasks is prioritized and high priority tasks are moved to the planning stage.

Plan Planning is done by a maintenance planner. For each task, the planner needs estimates for the following types of questions: How long will the job take? How much and what types of labor will be required? What parts and materials will be required, and are they on hand? What are the costs? What tools, equipment, or other resources, including external contractors, will be required? What permits will be required? What are the job hazards, and how will they be managed? Many tasks, such as inspections, periodic condition monitoring, and tasks with a safe work procedure and bill of materials, need limited planner input, but others, such as major asset shutdowns, need considerable input from the planning team. Ideally planning happens some weeks before the time period in which work is due to be executed as part of a well-regulated planning cycle. Once all the information is gathered a work order is planned and it can be scheduled.

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

Schedule Scheduling is the temporal organization of tasks for execution. It is a complex optimization problem with constraints such as the number of maintenance technicians available, limited ancillary equipment such as cranes, operational requirements limiting access to equipment, and system connectivity meaning that some work cannot occur at the same time as others. In addition, priority work must be balanced with preventive work so that preventive work does not fall behind over time.

Execute Considerable investment is incurred prior to execution due to the resources involved in planning and scheduling. Value from this prior work is added when the proposed maintenance work is executed by maintenance technicians through repair and replacement tasks to restore the required functionality of the assets. Good quality maintenance work restores the asset function to some required level or function, either as-good-as-new or some level between that and the current state. Poor quality maintenance work or work that is unnecessary can destroy value by introducing defects and cost money for little gain.

Complete When the technician has completed an assigned task, an important but often overlooked step is to capture data about the maintenance work with the as-found and as-left condition of the asset. This is documented on the work order, reviewed by the maintenance planner who is responsible for closing the work order, and stored in the CMMS.

Consequently, structured work refers to work that follows this entire maintenance management workflow. Structured work is planned and scheduled in longer time scales (normally planning sessions happen once a week and provide a time in the future to execute the maintenance). Unstructured work is often referred to as "reactive work", as these jobs result from failures that are identified by asset operators and executed immediately. These unstructured jobs are still completed and analyzed but do not pass through the formal planning and scheduling stages. Because unstructured work is executed immediately, it often results in other structured jobs associated with preventive and condition-based strategies to be rescheduled.

While these activities are the focus of maintenance reliability experts, this structure makes it difficult to discuss the human role in maintenance. The human actors within this workflow perform different tasks dependent on the situation (i.e., unstructured vs structured work). These roles and the responsibilities are described in Table 1.[1] However, while the person performing the task may change (e.g., a planner calculates time estimates for a job in structured work, whereas a technician might calculate time estimates on the fly for unstructured work), the tasks themselves largely remain the same. Regardless of context or situation, a human must 1. Prepare for the Maintenance Job, 2. Perform the Maintenance Job, and lastly 3. Discover Maintenance Needs. This distinction highlights how personnel actually perform each task and the types of errors that might occur in doing so, with a subsequent mapping from the task performed to the corresponding activity in the maintenance workflow for both structured and unstructured work. This task-based analysis is necessary due to the issues that still exist in manufacturing maintenance practice.

1.4 Issues in Practice

The SMRP Best Practices Committee suggests a distribution of maintenance work types as follows: for all executed maintenance work hours, 10 % to 15 % should be on improvement and modification work, 30 % on structured work - split between 15 % on predictive/condition-based work and 15 % on preventive work. Corrective maintenance hours derived from structured work should be 50 %, 15 % from preventive maintenance inspections and 35 % from predictive maintenance inspections. Only 5 % should be associated with corrective maintenance from unstructured work with a buffer of 5 to 10 % for other work. [12] In practice many manufacturing operations do not achieve these levels.

Small-to-medium sized enterprises (SMEs) still mainly employ a mixture of unstructured and structured maintenance strategies [18]. Once again, it is important to note that manufacturers often refer to corrective work as only unstructured, when in fact not all corrective work is unstructured work. Larger companies are employing preventive maintenance strategies, but unplanned maintenance jobs are still frequent [18]. Alsyouf [19] found that in Swedish manufacturing firms, 50 % of maintenance time was spent on planned tasks, 37 % on unplanned tasks, and 13 % for planning the maintenance tasks. Even though preventive maintenance strategies are more prevalent in larger companies, these maintenance jobs are not always performed correctly. It is estimated that one-third of maintenance jobs are improperly done or unnecessary [20]. Another study mentions that preventive maintenance is estimated to be applied too frequently in 50 % of all cases in manufacturing [21].

So why are so many SMEs employing mainly reactive unstructured maintenance strategies? Why are larger manufacturers still dealing with unstructured maintenance and often incorrectly performing preventive maintenance procedures? In a survey to manufacturers, the main barriers to adopting advanced maintenance strategies were cost (92 % of respondents), technology support (69 %), and human resources (62 %) [18]. This illustrates the need to help manufacturers find the most cost-effective path toward balancing technology solutions with human-driven tasks to improve maintenance procedure and reduce unplanned work. To effectively achieve such a paradigm, it is necessary to

[1] Different domains often use different terminologies for those roles. At smaller organizations, certain roles might be combined, such as a planner and scheduler or an operator and technician; however, for the purposes of this paper, we describe the roles as different people.

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

TABLE 1. Personnel in Maintenance

Job Title	Description of Responsibilities
Operator	Operates machines or monitors automated machines – can be responsible for one machine or multiple machines depending on the size of the organization and level of automation.
Technician	Used here to refer to the person performing minor maintenance jobs, for example routine inspections. These jobs can be done by both operators and maintainers.
Planner	Estimates time, cost, resources and documents for maintenance jobs, purchases parts and contracts.
Scheduler	Coordinate all planned jobs for a specific period into a realizable schedule.
Analyst	Analyzes and models data about equipment, operational and maintenance management performance.
Maintainer	A trade-qualified technician competent to perform tasks in their area of expertise.
Engineer	A degree-qualified individual who provides technical support for front-line staff such as operators and maintainers.

examine tasks within the maintenance management process to identify how to implement new technologies effectively by accounting for human knowledge and expertise.

NOMENCLATURE

AI Artificial Intelligence
AR Augmented Reality
CMMS Computerized Maintenance Management System
DES Discrete Event Simulation
ERP Enterprise Resource Planning
ETL Extract, Transform, Load
FMEA Failure Modes and Effects Analysis
HMI Human Machine Interface
IDEF Integrated Computer Aided Manufacturing (ICAM) Definition for Function Modeling
IoT Internet of Things
KB Knowledge-Based
MES Manufacturing Execution System
ML Machine Learning
MTBF Mean Time Between Failures
MTTR Mean Time To Repair
MWO Maintenance Work Order
NLP Natural Language Processing
OEM Original Equipment Manufacturer
RB Rule-Based
RCM Reliability Centered Maintenance
ROI Return on Investment
SB Skill-Based
SM Smart Manufacturing
SME Small-to-Medium Enterprise
SMRP Society of Maintenance and Reliability Professionals
SOP Standard Operating Procedure
SRK Skill-, Rule-, Knowledge-Based
SWP Safe Work Procedure
TPM Total Productive Maintenance
VR Virtual Reality

2 HUMAN FACTORS AND THE MAINTENANCE WORKFLOW

Incorporating a focus on human interaction with complex systems by applying human factors principles is not a new idea, and is rapidly gaining traction in sectors where implementation of new systems carries significant overhead, whether finiancially or culturally. In a 2011 report, the U.S. Department of Defense published a Human Systems Integration (HSI) Plan, [22] beginning with the following overview:

> *The human and ever increasingly complex defense systems are inextricably linked. [...] High levels of human effectiveness are typically required for a system to achieve its desired effectiveness. The synergistic interaction between the human and the system is key to attaining improvements total system performance and minimizing total ownership costs. Therefore, to realize the full and intended potential that complex systems offer, the Department must apply continuous and rigorous approaches to HSI to ensure that the human capabilities are addressed throughout every aspect of system acquisition [...] In summary, this means that the human in acquisition programs is given equal treatment to hardware and software.*

To accomplish this, human factors engineers will review functions and tasks within a system, which at their most basic assign responsibility of some activity to personnel, automated systems, or some combination thereof [23]. The primary goal of defining these tasks is to better understand not only the specific roles of personnel, but also how these will shift under implementation of proposed changes to the system.

Defining the role of human actors within a maintenance workflow has already been a core, if controversial, topic of in-

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

terest under the existing theoretical frameworks. Maintenance practices in the manufacturing sector center on the importance of individual authority versus the needs for centralized planning and scheduling of maintenance tasks. For example, the ideas of TPM focus on high levels of individual ownership of the asset by the operator with responsibility to adjust and maintain the unit. This will be largely an undocumented process with some work being done at the discretion of the operator to optimize their asset's performance. In this approach, the operator is empowered to take responsibility over the domain. This contrasts with the maintenance management view in which maintenance is centralized and the aim is to minimize costs across all equipment and resources and only to touch equipment if the work has been prioritized, documented, planned, and scheduled. With the introduction of more automation requiring individuals to assume responsibility for larger segments of the operation and with many highly-knowledgeable operators aging out of the workforce, the more centralized approach is gaining traction. However, the need for the human knowledge and expertise is greater than ever before and needs to be factored into the management process so as to optimize their contributions.

In systems engineering the incorporation of a human in any automated system is often *designed* as a fail-safe mechanism. Designing for all possibilities and failure modes is an impossibility, so designers and maintainers exploit one of humanity's greatest strengths: our ability to problem solve in unfamiliar situations and environments. However, this ability to reason from first principles has a high cognitive load, so humans try (where possible) to use rules and heuristics — mental shortcuts — to relax decision making and reduce the process of forming justifications into more habitual, routine tasks. Certainly, heuristics can be informed by prior observations of performance patterns and the success of previous solutions or approaches, but heuristics do not always work well to anticipate or mitigate failure: the human performing a task is left without enough context to recover from where the heuristic left off, or to estimate risks under unknown system behaviors or personal biases. Ironically, these situations tend to arise more often as systems become more automated—failure contexts become more complex and observations of particular situations become increasingly rare. The implementation of technologies, while intended to support technicians, will also require them to learn new ways of working. It will take time to build new sets of heuristics for each scenario. Digitization of equipment, for example, can decrease physical accessibility to manufacturing systems, along with altering the skill-set required to perform technical troubleshooting. The tension between a drive for automation to *compensate* for human error, and the necessity for humans to *compensate* for increasingly complex automated-system failures, should be dealt with up front by explicitly accounting for human failure modes that are causing the errors, in the original implementation plan. Orienting the function of emerging technologies in manufacturing maintenance around the causes of errors opens a path toward efficient and holistic implementation of those technologies.

This paper is not intended to serve as a sweeping guideline for implementing human factors, or for performing HSI, within maintenance in general—this would be far outside the scope for a single paper. Rather, we focus on specific pain points encountered in existing maintenance workflows, specifically in the context of human error before and after implementing some of the recently developing technologies in the space. We hope to provide initial guidance on augmenting specific functions/tasks within the maintenance workflow through certain types of technologies, based on how their strengths and weaknesses mesh with the strengths and weaknesses of critical personnel.

2.1 Human Factors Background

To analyze the maintenance management workflow, the role of the human in the maintenance paradigm must be understood. This paper uses the research by Jens Rasmussen and James Reason to provide a framework for estimating prime insertion points of new technology into the maintenance workflow. [24, 25] This framework provides guidance towards a hybridized maintenance workflow with both the human and technological systems working harmoniously. The framework centers around Skill-, Rule-, and Knowledge-Based error occurrences in the maintenance workflow.

Rasmussen introduced the Skill-, Rule-, Knowledge-Based Human Performance model in 1983. At the time when computers were becoming more mainstream, Rasmussen understood that the introduction of new digital technologies required "consistent models of human performance in routine task environments and during unfamiliar task conditions." This need for a human performance model ultimately led to the Skills-, Rules-, Knowledge-Based model of human behavior. Rasmussen proposed that human activity was a complex sequence of activities that depend on whether the activity was in a familiar or unfamiliar environment. He argued that, in a familiar environment, a human strives towards some high level goal through unconscious thinking based on similar situations. If the goal is not met, they use a set of "rules", which have perhaps been previously successful. In an unfamiliar environment, when proven rules are not available, a human makes different attempts – often in their head – towards a successful sequence to reach a goal.

Skill-Based Behavior A skill-based (SB) behavior takes place without conscious attention or control (e.g., tracking tasks in real time). A majority of the time, human activity can be considered a sequence of strictly SB actions or activities. SB behavior is an unconscious action implying difficulty or redundancy for a person to explain what information is required to complete the action.

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

Rule-Based Behavior When a sequence of actions is controlled by a rule or procedure derived from previous occasions, this is a rule-based (RB) behavior. The boundary between SB and RB behavior depends on the level of training and the attention of the person completing the task. Higher-level RB behavior is based on the human's explicit understanding of the problem and the rules used in accomplishing the task, while The "rules" in RB behavior can be derived empirically from previous attempts at solving a problem, communicated from another person's know-how, or may be prepared on occasion by conscious problem solving and planning. These rules are dependent on the knowledge of the environment.

Knowledge-Based Behavior When faced with an unfamiliar situation a human may need to rely on building new reference knowledge: this is knowledge-based (KB) behavior. A KB behavior involves explicitly formulating a goal based on an analysis of the environment and the aim(s) of the task. An individual develops different plans and tests them against the goal – either by trial and error or conceptually through understanding the environment and predicting the effects of the various plans – to determine the best course of action. This understanding requires mental models of the task and environment to predict the impact a specific plan might have on achieving the goal.

Error Classification While these different categories of human behavior are very useful in human reliability research, determining the appropriate category for individual tasks in a workflow is difficult in general. Thus, Reason [24] takes an error modeling approach toward the use of Rassmussen's skill, rule, and knowledge in his Generic Error Modeling System (GEMS). Rather than assigning a classification to each task, it is often more efficient to classify the error modes (which can occur while performing each task) into skill-, rule-, or knowledge-based errors. In the interest of using technology to cost-effectively address errors currently existing within a maintenance workflow, we focus implementation strategies around these errors explicitly. Figure 1 displays how this workflow is implemented in GEMS through the different levels of human performance, as well as providing some examples of various errors that can occur in the maintenance management workflow. The GEMS mapping of skill-, rule-, knowledge-based behavior onto errors enables an examination of activities within maintenance tasks by focusing on events when the system is *not* performing as desired — quite similar to system investigation through Failure Modes and Effects Analysis (FMEA). This discussion can help to determine the context-appropriate technologies that can be inserted into the maintenance workflow in a way that augments a maintenance practitioner's ability to successfully complete a task both efficiently and effectively.

1. Does the practitioner know that something is amiss while it is happening (has an attentional check occurred)?

No The errors involved will be SB level. Mitigations for these would help him/her perform the attentional check (notice the error), or make noticing at this part of the workflow unnecessary through anticipation.

Yes Problem is being investigated at a RB or KB level

2. Does the practitioner believe they have a way to solve the (noticed) problem?

Yes Errors will be RB level. The selected "rule" may not actually be appropriate, and mitigations should provide more (or better) sources of data and pattern discovery, e.g., sensor outfitting, machine learning models.

No They are actively searching for a new rule, making the relevant errors KB level. Context and causally sensitive models would be helpful to teach or suggest new solutions, like simulation, schedule optimization, or expert systems.

Using this model, the same failure in an activity may have distinct causes — understanding at which level the failure occurs can help to address it. An SB error stems from the inherent variability of human action with familiar tasks. Commonly referred to as slips and lapses, they generally occur without immediate recognition that something is wrong. It is only after an "attentional check" that one might notice something has gone awry, and begin applying some known rule or pattern that addresses this problem.

RB errors typically are the misclassification of situations, leading to the use of an inappropriate rule or incorrect recall of procedure. However, once the problem-solver realizes that none of their existing rules apply, he/she begins modeling the problem space (e.g., by analogy) and searching for context clues that relate the problem to past successful rules. KB errors arise from the incomplete or incorrect knowledge of the problem-solver and stem from situations that represent the highest cognitive demand. Reason indicates that this state will quickly revert to SB once a satisfactory solution is found, and that this is a primary cause of sub-optimal solutions. GEMS postulates humans behave in such a way as to minimize their cognitive load and that many errors are a result of this tendency. Once an error is recognized, the person will move to the next higher level of cognition to resolve the error; once resolved, they will quickly retreat to lower cognitive effort.

2.2 Error Mitigation Framework for Efficient Technology Adoption

This process of discovering a problem and the validity of the current solution strategies is the same process that will be applied at a management level when implementing new technologies into the maintenance workflow. Unless it can be demonstrated that 1) there might be a problem with the current situation (SB) and 2)

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

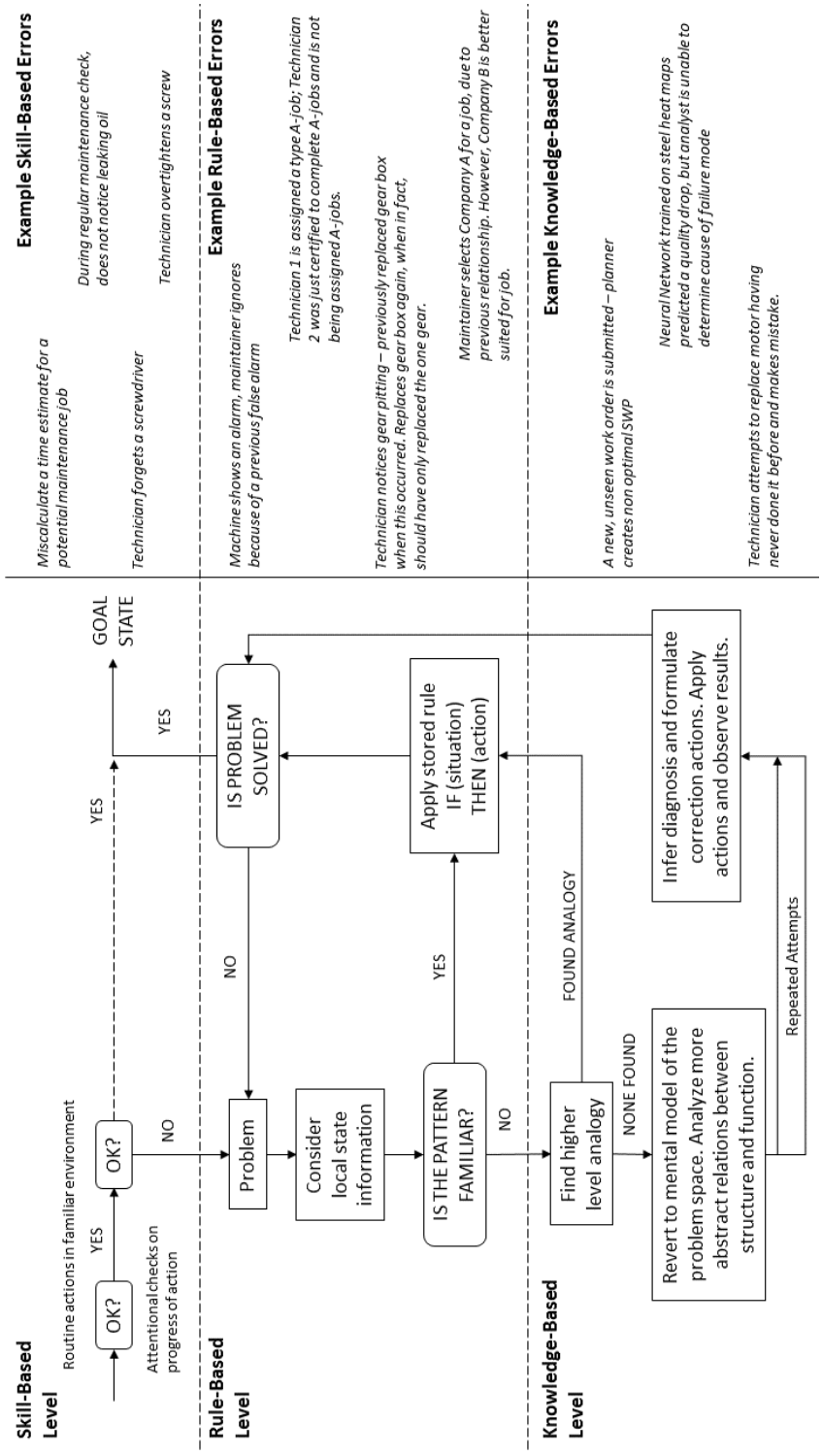


FIGURE 1. SRK Framework. Adapted from Reason [24].

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian. "Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing." Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

the solutions currently in place are sub-optimal (RB), practitioners in the existing maintenance workflow will be unlikely to use solutions that attempt to improve performance through suggestion of new modes of operation or behavior (KB). Knowing how to frame technology implementation in terms of these steps is key. Cultural momentum and the power of the status quo is consistently overlooked, and failing to understand or adapt to it is nearly always the primary cause for technology implementation failures [26]. While this may sound extreme, in the context of system maintainers, it makes sense: *if trusted personnel in the human-centric maintenance workflow do not believe that there is a problem, or do not think their solution is insufficient, the possible performance of new technology will not come to fruition—no matter the expenditure that went into implementation.*

This paper does not exhaustively enumerate all potential errors, their probabilities, or the factors that affect them. Starting on that path would require a more sophisticated Human Reliability Analysis (HRA). [2] We leave this worthwhile task for future work. Instead, we illustrate some common errors that can occur in the maintenance workflow and how GEMS can be used to highlight common opportunities and pitfalls when implementing technologies meant to assist the maintenance management process. Sequentially progressing through SB, RB, then KB errors and deciding the risk and mitigation possibilities of each through available resources provides valuable guidance, especially when choosing a starting point can be difficult in the face of capital or personnel costs. Importantly, we

(a) classify example errors that commonly occur during maintenance tasks, and
(b) discuss how the features of each task tend toward more or less errors of a given type.

Going through a similar exercise prior to selecting or implementing new hardware or software systems will assist in matching use-cases, as well as mitigating consequential errors effectively, since different technologies are designed to address widely different error types.

A tiered error-mitigation strategy, based on patterns observed in literature and industrial application, structures our discussion of inserting advanced technologies into a maintenance workflow. Improvement opportunities are aligned to the consequences of the problems they address, balanced with feasibility of implementation in terms of cost, logistics, and organizational maturity. In this strategy, errors are addressed based on their cognitive load.

The discussion that follows centers around the different errors in the maintenance workflow and the human actor who commits them. It is important to note that these errors are committed by a variety of roles, and not necessarily by just the technician (as is often thought by maintenance managers). The errors presented are often discussed on an individual basis (e.g., one technician does not notice an alarm); however, manufacturers must view these errors at a systematic level to understand the true "pain points" in their factory. A single technician not noticing an alarm is not typically high risk, but having a majority of technicians systematically miss alarms is a larger, more important issue to recognize. As such, while it is tempting to focus on individual actors as problem sources, better guidance is needed that assists manufacturers in tracking and estimating errors across the entire factory.

In the following sections, applicable research and technologies are discussed that apply to the tasks in the maintenance workflow. Examples of errors are described for each task: 1. Prepare for Maintenance (Section 3), 2. Perform Maintenance (Section 4), and 3. Determine Maintenance Needs (Section 5), and errors are mapped to the sub-tasks in Tables 2, 3, 4. Each table maps sub-tasks (Column 1), to example errors (Column 2) and their corresponding GEMS classification (Column 3). The errors and mitigations as presented are intended to be exemplary of common errors practitioners will encounter in the maintenance workflow.

3 PREPARE FOR MAINTENANCE

Prepare for Maintenance involves a number of actions to enable execution of maintenance work. The tasks performed by the human actors in maintenance are largely the same, but are performed in different stages of the maintenance management workflow and are performed by different people depending on structured versus unstructured work. For structured maintenance events, the required tasks are prepared by a maintenance planner/scheduler over a period of days, weeks or months. During unstructured maintenance events, the jobs and required actions are identified, often by an operator or technician while in the field, and usually under time constraints as the component may have already failed. A number of research efforts center around the prepare for maintenance task, as discussed in the following subsection.

3.1 Applicable Research & Technologies

Maintenance preparation is a very human-centric operation relying on tacit knowledge of how similar jobs have been planned in the past, what has worked well, and what has not. Various efforts have codified this knowledge using safe work procedures, bills of materials, and post-work reviews [10, 9]. As the balance of work to plan moves from corrective to preventive and predictive work, the opportunity for semi-automation of the maintenance planning process will increase.

Despite the considerable academic focus on maintenance scheduling under the umbrella of maintenance optimization, the

[2] A combination of human factors task analyses and systems engineering FMEA. See HRA frameworks in Kirwan, Gertman, and Hollnagel [27, 28, 29]

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

levels of transaction automation and the use of simulation models in this process are small. Maintenance optimization uses mathematical models to find either the optimum balance between costs and benefits of maintenance or the most appropriate interval or moment to execute maintenance. An overview of the maintenance modeling approaches and examples of their applications are available in Dekker (1996 and 1998), Marquez, and Jardine [30, 31, 32, 33]. Both engineers and mathematicians have contributed to the area. Due to the complexity of these models they have not been easy to apply to real world manufacturing systems in practice [30].

Maintenance simulation models are classified in a number of ways. First, are they for planning or scheduling? The vast majority of optimization models in the literature address scheduling. Secondly, at what level is the maintenance decision being taken: organizational, plant, system, unit, or component? A consequence of the level consideration is that decisions at higher levels need to take all lower levels into account. Many different types of dependencies must be considered and these can only be accounted for in a simplified and often inaccurate way [31]. Finally the model needs to consider time scale. Is the model to support a decision with impact in the near or long term? Are we thinking about the next schedule period or something that could have long-term, but deferred impact, on the life of the asset?

Discrete event simulation is widely used to model maintenance systems and the uptake of new optimization methods, such as genetic algorithms, has been rapid. However, a review by Alragbhi [34] found only a few real life case studies were published and the academic cases that dominate the literature, such as a single machine producing a single product, are oversimplified and do not reflect the complexity and interactions of real systems.

Scheduling practice differs greatly depending on the type of system. Scheduling practices in manufacturing differ depending on whether the operation is a batch process or continuous process and on the availability of buffers. The presence of buffers in the system allows for a more flexible approach for minor maintenance and for opportunistic maintenance to occur [35, 36, 37, 38]. Examples of work on maintenance in a manufacturing context mainly have focused on management of preventative maintenance activities [39, 40] or maintenance resources [41] to optimize manufacturing system performance.

Practical and applicable models are needed that derive a set of optimized maintenance schedules offering a range of tradeoffs across the objectives from which managers can select for their immediate needs. The modeling system needs to be able to adjust schedules on a real-time basis as circumstances and/or priorities change. Maintenance optimization is a complex problem, with multiple possible objectives such as system reliability, cost, availability and various combinations of these (many plants easily involve over 100,000 periodic activities). The complexity of the optimization has often precluded the use of decision guidance systems in real-time under current practices. Emerging technological advances are enabling better support in these systems, however new solvers are required to develop and solve the proposed models. Too often in the past engineers have focused on optimizing a particular asset or subsystem, where the complexity is more manageable, rather than considering the entire maintenance management system.

Digital twins are an emergent focus for many in manufacturing and are an integral part of Smart Manufacturing [42, 43, 44]. A digital twin is a digital model of the asset system. It is constructed using digital information of the physical asset and its environment and can be continuously updated from sensor data. This should enable better planning, prediction and simulation of future outcomes.

Despite the widespread use of discrete event simulation models, commercial and research interest in the potential of agent-based simulation approaches is increasing, particularly when organizational and human factors need to be incorporated [45, 46].

As discussed earlier, it is not simple to incorporate these technologies seamlessly into the maintenance management workflow. By decomposing the perform maintenance task into sub-tasks, we can better analyze the types of errors that occur and the potential mitigations. These sub-tasks include: 1) Identification – considering steps necessary in the maintenance execution process, 2) Planning – determining the required resources to perform the jobs, and 3) Scheduling – determining the schedule, when the job will be performed, and in what sequence with other jobs. The typical errors for this stage are described in Table 2.

3.2 Identification Task Errors

The identification of work occurs through the structured processes and also during reactive work as described in Section 2. In the latter case the maintainer must identify the work to be done when he/she gets to the failed asset. Similar human processes are involved in both examples of 'identification'. Some examples of errors that occur during the identification task are below.

SB *A CBM technician fails to notice the vibration sensor is not adhered properly so the data collected is wrong.* (Assess Sub-task)

RB *A CBM technician identifies a high peak in vibration when collecting data on Pump 1 but the source is actually the adjacent pump.* (Assess Sub-task)

KB *CBM monitoring technician generates a work order that machine X1's "lead-screw's vibration is high". This failure mode has not previously been seen and there are no visible symptoms. The proposed work is subsequently overridden by planner who believes the analysis is inaccurate.* (Assess Sub-task)

Traditionally the approach to dealing with SB errors on the

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

TABLE 2. Prepare for Maintenance Tasks: Errors and Mitigations

Sub-task	Example-Error Description	Error Type		
Identification Tasks				
React	Failure not noticed	SB		
	Failure ignored		RB	
Anticipate	Incorrect interval for replacement		RB	
Assess	Incorrect condition for replacement		RB	
	Operator misses condition	SB		
	Misunderstand condition calculation			KB
Planning Tasks				
Estimate	Error in calculation	SB		
	Match task to wrong SWP		RB	
	Estimate from wrong sources of data or experience		RB	
Procure	Misevaluate available resources (when to purchase parts/use in-house)		RB	
	Mismatch contractors or pick the wrong parts		RB	
Prepare/Document	Miss/overlook a necessary document	SB		
	Include or require wrong documents		RB	
	Proposed procedure is non-optimal			KB
	Ignore previous feedback from execution team			KB
Scheduling Tasks				
Prioritize	Poorly track resource availability	SB		
	Misevaluate interrelationships between different maintenance tasks			KB
Assign	Poor match of skills of technicians to the job		RB	
	Lack of communication between executing team and scheduler			KB

plant floor is through increased surveillance with the use of sensors, process control and alarms. These techniques give more than one person the opportunity to identify the fault. Another common approach has been the development of checklists. However, these checklists can promote mindless completion of forms – even if the form is incorrect or incomplete – for the sake of just completing the form because they are told to complete it, instead of mindfully completing the forms to properly and accurately document the activity. A proliferation of unqualified checklists can also create issues and, where they are useful, should be replaced by centrally developed and version controlled maintenance procedures. Currently, few technical solutions address when the planner is dealing with paperwork when many distractions and other calls on his/her time exist. Improved supervision, workload management and team support are often key to improving concentration and mindful execution of routine work [47, 48].

RB and KB errors (See Table 2) in identification of work often result from different mental models of the failure or its consequence between parties involved. As in the illustration above, the technician assumes the vibration data is from Pump 1, according to his/her previous experiences, and so assumes the data is correct even though it is incorrect.

These errors can be mitigated through investment in digitization of the prioritization and approval processes for the planner or through improved training (for the maintainers). Ensuring that the initial SB errors are mitigated first where possible, enables the higher-level RB and KB errors to be addressed through these more advanced approaches.

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

3.3 Planning Task Errors

Planning involves estimation of necessary resources, time, cost for each job based on historical organizational data, rules and practices and the experience of the human planner. Parts must be procured and maintainers and tools selected, and documents are prepared to assist in execution. The tasks within Planning are similar for both structured and unstructured work. However, for structured work the tasks are performed on a longer time scale and by a dedicated planner. This contrasts with the shorter time scale (often right when a failure occurs) associated with the unstructured work done by an operator or technician. A summary of common errors and their classifications are located in Table 2 and some examples are discussed below.

SB *When planning a rebuild on Machine X the maintenance planner miscalculates the time estimate for job 1.* (Estimate Sub-task)

RB *The planner orders all the same parts as used in the last Machine X rebuild rather than considering the work specifically identified for this rebuild.* (Estimate Sub-task)

RB *The new maintenance planner contracts Company A for Machine X rebuild, because of a past relationship, but Company B should also have been considered.* (Procure Sub-task)

KB *The planner miscalculates the downtime required for Machine X rebuild by failing to take account of resource constraints.* (Prepare/Document Sub-task)

SB errors during this task, such as forgetting a necessary document or making a slip during an estimate calculation can be aided by centrally managed and controlled procedures that are easy to use. Many of the errors during this task are RB errors that involve matching an aspect of the work order to some necessary document or resource. These type of errors could be well suited for machine learning solutions because these algorithms can learn the important features of the maintenance task and match to the correct previous solutions to provide estimates of resources, time, cost, etc.; however, these solutions can be difficult to implement because of the way in which data about the tasks is stored (natural language) and because of the variety of different contexts in which the same task can be executed. If these SB errors are dominant, investment in the search-based solutions enabling planners to locate information on previous similar tasks can assist.

3.4 Scheduling Task Errors

Once the tasks are planned, they must be scheduled for a specific time. For structured work, this task is completed by a scheduler. However, it can be done in the field by a technician or operator when a failure occurs and an unstructured job is initiated. Coordinating the scheduling of many machines, people, parts, contractors and production equipment requires consideration of many permutations for optimal solutions. This complexity can potentially lead to a number of higher level errors such as those seen in Table 2 and discussed below:

SB *Maintenance scheduler forgets that Team A has a safety Day and schedules work when they are not available.* (Prioritize Sub-task)

RB *Technician 1 is chosen for an A-type job, as always; Technician 2 has recently been A-certified, and is not being assigned A-type jobs.* (Assign Sub-task)

KB *The scheduler reuses a schedule for rebuild of Machine A event though analysis of the the last time this was done resulted in a 50% time overrun.* (Assign Sub-task)

Given the complex nature of scheduling, a high incident of KB errors are likely to occur. These types of KB errors can benefit from investment in project planning software and scheduling and optimization models. SB errors, when a scheduler has a lapse determining availability of a technician or asset can be mitigated through scripts to assist in availability calculations from calendars. RB errors that occur in matching a work order to an appropriate technician can benefit from the analysis and modeling solutions discussed in the above section on Planning tasks. Scheduling is one of the more difficult tasks to provide easy-to-implement solutions; however, it can benefit from planning tools that enable a variety of schedules to be tested and various constraints to be incorporated.

4 PERFORM MAINTENANCE

The Perform Maintenance stage consists of executing the maintenance actions, and recording the necessary information about the job. The majority of tasks are similar whether the work is structured or unstructured. The research efforts in this space are described in the following subsection.

4.1 Applicable Research & Technologies

The perform maintenance task includes the maintainer documenting the as-found and as-left conditions of the equipment as well as the work that was done. The steps in executing maintenance work have changed very little in the last 40 years. Many of the same tools and processes are used. There have been advances in support tools, for example laser alignment to replace dial indicators, auto-lubers, and greater use of digital interfaces to support troubleshooting for electrical/electronic equipment but the nature of the way work is executed today would be familiar to many retired technicians.

Each time a maintainer or technician interacts with equipment, he or she expands their own expertise of the asset and captures a textual description of observations of the asset and records what was done, when, and how. The lack of correct and complete data in work order records to support analysis is widely ac-

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

knowledged. Recent work to better understand factors that affect data quality of maintenance work orders include [49, 48]. Work orders typically contain unstructured text with jargon, abbreviations, and incomplete data. Primary interests for analysis are information to establish the as-found condition, the causality of failure including the failure mechanism, and a description of the maintenance work executed and parts used. This data is often in the work order texts, but it is not extracted in a machine-readable way. As a result maintenance staff rely heavily on personal expertise, word-of-mouth, and ad-hoc data exchange, consulting the records when these other methods fail.

Research from several different academic perspectives has been conducted on the execution of maintenance work; however, these types of studies have seldom translated into meaningful change on the maintenance shop floor. Human factors specialists have looked at how maintainers interact with assets [50]. The impact of human error on maintenance outcomes has been of significant interest [51] and spurred attention from other organizational psychologists in exploring how culture affects motivation and the execution of quality work and consistency in following procedures [52, 47]. Considerable interest exists now in the potential for mobile technologies such as assisted reality and GPS tracking to better understand and support maintainers in the field both in the execution of their work and in how data about the work is collected [53]. The latter is of vital interest to engineers as a maintainer's observations on the as-found condition of an asset can be vital input to validating condition-based work orders.

The explosion of current technology dealing with multimodal data sources is particularly relevant to maintenance management. The information about asset condition, failure cause, and maintenance work extends beyond what is captured using language in written work order records. While work orders are central to maintenance processes, maintainers communicate with each other and others using a wide variety of media, including photos, videos, emails, text messages and phones, in addition to other resources such as sensor data. Support systems are emerging to provide access to critical information about maintenance issues from disparate sources. Given the emergence of alternate ways of collecting data from maintainers with mobile devices containing cameras and audio to sensor data from machines, methods to efficiently process and synthesize these different modes of data capture to provide asset health status assessment are needed. Assisted and augmented reality (AR) head-set systems are emerging into the market that provide maintainers with access to audio and visual support in the field and the ability to look at drawings and other relevant information in a head set visor [54, 55, 56].

Technical developments are needed to enable maintainers to efficiently capture, retrieve, absorb, process and exchange knowledge about equipment and maintenance work. One of the most exciting recent developments is in natural language processing to enable work order texts to be read and analyzed more efficiently by computers. Examples of recent work in this area include [57, 58, 59, 60]. This work is complemented by developments in semantic knowledge representation technologies to capture data and contextual relationships between data. This will support the development of inference engines capable of performing basic reasoning over maintenance operations enabling better decision support and improved quality control.

Another notable development is the emergence of ontologies for maintenance that are aimed at addressing different needs such as data integration, semantic interoperability, and decision support in maintenance. For example, several ontological approaches have been proposed to overcome the problems of heterogeneity and inconsistency in maintenance records through semantic data annotation and integration [61, 62]. When formal ontologies are used for annotating vast bodies of data, this data can be more easily retrieved, integrated and summarized. Also, the annotated data can easily be exploited for purposes of semantic reasoning. In a recent initiative, referred to as the Industrial Ontologies Foundry (IOF), an international network of ontology developers are working towards developing a set of modular, public, and reusable ontologies in multiple industrial domains [63, 64]. Their work includes a reference ontology for maintenance.

To discuss how the technologies within this stage can be implemented in the maintenance workflow, the perform maintenance task is decomposed into the following sub-tasks: 1) Assessing and Diagnosing, 2) Executing the Maintenance Action, and 3) Completing and Recording the Action. The set of typical errors for this stage is in Table 3.

4.2 Assessment and Diagnostic Task Errors

The Assessment and Diagnostic Tasks depend on the type of work required, such as assessing the equipment condition to see if condition of the asset is as expected from the work order and if the task described in the work order is appropriate. These tasks rely heavily on the tacit knowledge of the maintainer, whether heuristics or rules-of-thumb, that can be applied in uncertain or developing circumstances. One way to think of this is if an assessment does not match the assigned work, similar diagnostic and assessment tasks are required as in the reactive identification tasks, discussed above. This means that many of the SB and RB errors mentioned previously (specifically for unstructured work) apply to this task as well.

In supporting these tasks maintenance technicians may face specific challenges. If they are in the field (this may be remote from the maintenance shop such as on the factory floor), they can be isolated from reference material or knowledge bases and their team and supervisors. In addition often ergonomic constraints exist — e.g., using a touchscreen is difficult with gloves on that makes ready access to digital support tools challenging. In ad-

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

TABLE 3. Perform Maintenance Tasks: Errors and Mitigations

Sub-task	Example-Error Description	Error Type		
Assessment and Diagnostic Tasks				
Assess	Overlook symptoms that indicate poor equipment health	SB		
	Incorrect condition features used in assessment		RB	
Compare	Incorrectly assume validity of assigned work		RB	
Diagnose	Unfamiliar symptoms lead to incorrect fault diagnosis			KB
	Incorrect diagnostic conclusion due to lack of experience			KB
Execute Maintenance Action Tasks				
Perform	Forget necessary tools needed to complete job	SB		
	Lapse in execution quality due to focus constraints	SB		
Triage	Mistake nature of work for similar type having distinct solutions		RB	
	Attempt execution without requisite experience, tools or supervision			KB
Completion and Recording Tasks				
Recall	Technician does not remember significant symptoms	SB		
	Recalled features are not relevant to analysis		RB	
Record	Work performed is entered incorrectly, or schema structure is incomplete		RB	
	Technician gives up searching prior to finding appropriate problem-code		RB	
	Technician misunderstands or is unaware use-case and functionality of the data structure (e.g. controlled-vocabulary)			KB

dition, digital support tools need to be rugged to survive dust, water and unsecured work places that can be present in maintenance situations. This isolation from easy-to-access reference material differentiates this step as having a high concentration of possible KB errors, for example:

- **SB** *There is a noise in a pump-motor unit, the technician notices the noise as assesses it as a potential failure but gets side-tracked and fails to report it.* (Assess Sub-task)
- **RB** *The engineer investigates the noise in the pump-motor unit in the field but decides it is 'normal' when it is not.* (Assess Sub-task)
- **KB** *The vibration analyst diagnoses the pump has a bearing failure due to lack of lubrication but the cause was a seal failure.* (Diagnose Sub-task)

The sources for error during Assessment and Diagnostic tasks that have significant impact are less often slips or lapses in memory (SB errors), but rather stem from the complexity in diagnosis of the cause of machine failures or sub-optimal performance. Machines of the same make and model can be at different life stages, have experienced different operating profiles and maintenance events. This means that all machines are subtly different and hence diagnostic rules that should work on one machine, do not always work on another. This situation results in RB and KB errors during execution. The need to address unfamiliar situations in the field, often in a remote location and without immediate access to knowledge bases, compounds the need for more sophisticated approaches toward technological enhancement of agents executing maintenance. Rapid advances have been made in assisted reality glasses and headsets for diagnosis. These tools need to be supported by trained people and new business processes.

4.3 Maintenance Execution Task Errors

The Execute Maintenance task occurs when the maintenance action is explicitly performed. Work can be a routine job (performed regularly) or non-routine, involving new steps that may not be familiar. The Execute task is also highly human-dependent. Routine work includes many SB errors, while non-routine work involves more RB and KB errors, as described below.

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian. "Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing." Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

SB *Technician 1 forgets to loosen the motor when installing new v-belts.* (Perform Sub-task)

RB *Technician knowingly replaces only one of the v-belts rather than the whole set as he/she had done this last time and there has been no adverse repercussions.* (Triage Sub-task)

KB *The technician knows that an adjustment in alignment needs to be made for thermal growth but cannot remember the rules and formula.* (Triage Sub-task)

The errors out of Execute Maintenance are dependent on the type of work. Issues like forgetting a set of tools before reaching a job location, or accidentally forgetting to loosen the motor while in a hurry, are not necessarily able to be mitigated through direct automation. Rather, these tend to ease over time on an individual level with experience on the shop-floor and better planning. Obvious aides like digital assistants may be additions to speed up this process, however, automation systems are not capable of replacing a human at the skills-level in this manual, tacit, dexterity-intensive task. RB and KB errors in this task stem from lack of appropriate experience of the technician. One approach to mitigating these inexperience errors, especially where effective rules *do exist* in the expertise of senior staff, is a buddy system [65]. Such a system is increasingly being digitally augmented through the use of assisted reality as described in Section 4.2 or remote support systems which a technician to bring in an expert virtually.

For the KB errors, training is always useful, though it is impossible to train for every occurrence. Knowledge-bases tend to be of limited use here, since ergonomic constraints (like gloves, ambient/background noise, etc) make interfacing with traditional digital systems — or even documentation — rather difficult. However, recent developments in assisted- and augmented- reality displays (AR/VR) can bypass this ergonomic problem, especially in preparation for jobs on difficult or seldom accessed equipment. It is important to remember that these displays do not provide such functionality out-of-the-box, and several supporting technologies, like interconnected data storage and digital twin reference models, will need to be successfully adopted prior to reaping benefits from the continually decreasing cost of this exciting technology.

4.4 Completion and Recording Task Errors

When addressing the state of data recording in maintenance, regardless of sensor-outfitting or other types of data-streams, one goal of recording maintenance information is to capture the activities of a person *executing maintenance* — their ability to diagnose and solve problems. Recording this information requires the technician to recall features associated with the work order that distinguish it from other work orders. Once these features have been recalled, they must be recorded by translating into a format acceptable for predefined data structure required by the CMMS.

Recalling features poorly is typically a sign of unstructured work. Slips and lapses are more likely to occur in this recall phase, e.g.:

SB *Emergency MWO's 18 and 19 were executed yesterday, but work pressures meant the information was not recorded until today. Cannot recall which of MWO 18 or 19 was the seal replacement on Machine A3* (Recall Sub-task)

Structured work tends to reduce the likelihood of this type of failure mode: pre-documented assignments that have been planned and scheduled a priori will have associated documentation that assists or automates a significant amount of feature recall, leaving only the most relevant human actions to be input by hand. However, simply using a work order generated by a CMMS does not by any means guarantee high quality data. Translating data into a CMMS will necessitate higher-level cognitive engagement, and associated errors quickly transcend the skill-based level:

RB *Technicians have been asked to classify failures with specific names and codes to help with analysis but Technician A continues to use the names he/ she has used in the past instead.* (Record Sub-task)

KB *There are 5 fault codes and the technician struggles to find one that actually describes the fault, so selects the "miscellaneous" code but provides no further information.* (Record Sub-task)

Addressing these types of errors is more difficult. Solutions to the SB errors (e.g., standard MWO structure, pre-filled MWO forms, designated time-slots for data entry after every work order, etc.) could provide significant return on investment to improve data quality, and should be in place prior to addressing the higher-level RB and KB problems. Recommendation systems and user-interface design can be helpful in improving potential value of the recorded data. Statistical summaries of common themes in existing "miscellaneous" work order records, through the use of Natural Language Processing, is potentially useful, though care to include expert judgments must be taken when processing technical, domain-specific, short-hand-filled language [58, 60].

Recommendation systems could be applied during the completion stage to augment a technician's ability to rapidly sort his/her knowledge into the required format. [66] If sufficient effort has been made to create and maintain digital references for an entire line, real-time suggestions for recording related symptoms or components could provide a boost to both data-quality and the speed of experience-gain for the maintenance team. Given sufficient investment, these tools could provide additional input and context that assists technicians in creating the

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

rules and *knowledge* they need — combating the errors induced from often-dense, difficult-to-navigate user interfaces for selecting from a complex web of controlled vocabularies that so often occur in this space.

5 DISCOVER MAINTENANCE NEEDS

Discover Maintenance Needs tasks involve the use of software tools to create value from existing data, and inform the future workings of other tasks. These tasks are independent of structured versus unstructured work, but the tasks performed inform future structured work. Technology and research in this stage are described in the next subsection.

5.1 Applicable Research & Technologies

Discovering maintenance needs should be the product of a maintenance strategy informed by on-going analysis of asset condition, performance, and failures. Maintenance strategy is informed by an understanding of the function of the component, its failure behavior, and the consequence of loss of function as determined by a FMEA [67]. A Risk Priority Number is produced based on the likelihood, consequence, and detectability of each functional failure. Maintenance strategies are developed for the most critical functional failures using a Reliability-Centered Maintenance (RCM) or similar process [68, 69] and described in a variety of standards [1, 70]. Common names for these strategies are *design out* (or *improvement*), *predictive and condition-based, preventive, failure finding*, and *run-to-failure*. Condition-based strategies produce tasks to collect and analyze the performance or condition data (but not to do the work arising from the analysis). Run to failure strategies, employed when there is low consequence of failure and the cost necessary to prevent it exceeds the cost of the failure, result in corrective maintenance work. For RCM, the interested reader is referred to Rausand [68] and examples from infrastructure applications, such as electric power distribution systems [71, 72], maritime operations [73], and wind turbines [74]. Although RCM is widely used in defense, automobile, aerospace, and electronics for product design, there appears to be limited, well-cited literature, such as Tu and Jonsson [75, 76], on the use of RCM for the equipment used in manufacturing processes. A potential roadblock to the implementation of novel sensing and analytics opportunities is a manufacturing plant's lack of a well-framed and functioning maintenance strategy process.

The subject of analysis in maintenance work is vast and encompasses topics such as reliability analysis, health condition diagnostics and prognostics, predictive maintenance models, strategy selection models, maintenance performance, and spare parts modeling. Despite the growing number of papers published on these topics each year, the uptake of the various models by industry is low [77]. Much work, particularly in prognostics has been theoretical and restricted to a small number of models and failure modes. There are few published examples for manufacturing systems and, more generally, on systems exposed to a normal range of operating and business conditions [78]. Published models rarely examine their practical and theoretical limitations in sufficient detail to understand when and where the model should and should not be applied. Other issues include how to assess model performance and uncertainty quantification [79].

Although asset manufacturers and operators have used sensors and manual data collection for decades to collect health data on assets, developments towards IoT offer a new opportunity where data is transmitted from assets to the Cloud [80]. In this architecture, data for health estimation (e.g., condition monitoring, environmental condition data, previous maintenance work, and operating data) is more readily available for health monitoring and prognostic assessment to assist in identifying maintenance work. The cost and ease of sensor deployment also creates opportunity for more relevant data collection. This sharing of information across assets and platforms should enable the development of a system view and the flexibility to assess and manage existing and emerging risks [80].

The potential for these developments to impact manufacturing maintenance is widely evident. Asset-owning companies are implementing software platforms to better understand their maintenance needs especially for predictive maintenance applications. The market for software platforms to integrate data from multiple sources and support the use of this data in analytics and access to the results through web applications is dramatically growing. Examples include, but are not limited to, Dassault Systems 3DExperience [81], General Electrics Predix [82], PTC ThingWorx [83], Inductive Automation Ignition [84], and Siemens Mindsphere platforms [85]. Beyond the commercial market, free, and accessible open-source statistical and machine learning algorithm packages, online training, software platforms and visualization applications are making these capabilities accessible to manufacturers on a lower budget.

While these technologies expand the potential for better maintenance practices, they are not without complication. The growth in research in the application of the technologies is enormous but the path to wide-spread application has not yet been paved. Challenges exist with both identifying the opportunities for better prediction and with creating the infrastructure needed to get the right data at the right time.

There is no shortage of predictive maintenance models proposed in the scholarly literature with over 25,000 'predictive maintenance' papers listed in Google Scholar in 2018 alone. As a result of the growing choice of available diagnostic and prognostic models, a number of papers have been written to provide guidance on model selection, for example in Leep, Lee, Sikorska [86, 87, 78].

Machine learning technology is proving useful in this context. Machine learning models train on large amounts of data

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

to provide output predictions given new input from many large historical datasets (e.g., Neural Networks, Support Vector Machines, Bayesian Networks). Provided relevant and sufficiently-sized datasets, these data-driven models can be good at detecting and predicting poorly understood or poorly modeled system behavior without a strong dependency on the relevant physics or other dynamics. However, the nature of failure datasets creates particular challenges. Failures, particularly of critical equipment, are rare. Most equipment is replaced in whole or in part before the end of life. As a result, failure datasets are unbalanced and sparse. In addition, for reliable analytics the datasets need to be assigned meaning, or *labeled*, which can be an onerous task. Without this labeling, the ground truth for validation often does not exist. Furthermore, condition monitoring data, when available, is often collected on assets using different methods at different time intervals which complicates the analysis process. Poor quality data results in greater complexity of the analytic models that at best muddies inference and at worst misleads inference and produces persistent prediction bias. These contextual issues, if not dealt with rigorously in model selection and validation practice, lead to poor performance and a loss of trust by decision makers.

Another challenge in deploying analytics is that each data-driven model is developed for a specific application, resulting in the need for a plethora of models depending on the scale and complexity of the manufacturing system. At a minimum a prognostic or diagnostic model needs model selection justification, validation, application limitations, and uncertainty quantification [77]. Developing a model for each dominant failure mode involves significant time and cost, which is increased by maintenance and validation of the model as the asset ages or operating conditions change. Considerable opportunities exist to develop new processes, platforms, and standards –an ecosystem– to support these models enabling them to be more widely adopted.

To aid in technology insertion, the determine maintenance needs task is broken into sub-tasks: 1) Data Extraction, Transformation, Loading (ETL) for organizing data and 2) Modeling and Analysis. The tasks are mainly performed by engineers and data scientists, requiring system architecture knowledge, expert elicitation, and mathematical or physical modeling assumptions. The common errors for this stage are described in Table 4.

5.2 Data Extraction, Transformation, and Loading Task Errors

Data does not exist in a vacuum, and cannot provide value without intermediary steps. Collecting, storing, processing, and serving data to analysis tools are all core parts of data engineering and are relevant to MWO data. The tasks required of data analysts are typically organized as ETL. Extraction refers to getting data from relevant data sources, such as machines on the shop floor. Transformation is the act of preparing the data, such as cleaning, data type selection, or feature engineering. Loading is the process of sending the data to the another system for modeling and analysis. Because the goal of a properly implemented ETL system is the automation of data transfer and organization, sources for KB errors are typically few, occurring around misuse of software or functionality. Rather, key errors possible in these tasks are largely SB and RB:

- **SB** *Data tables do not record units for the sensor readings.* (Transform Sub-task)
- **SB** *Data engineer merges two tables using the wrong join process.* (Transform Sub-task)
- **RB** *The data analyst uses an Excel spreadsheet from a colleague without checking if cells are updating properly.* (Load Sub-task)
- **KB** *Reliability engineer when loading data puts zeros into empty NA cells which skews all the subsequent analysis.* (Load Sub-task)

Lapses in design specification, like forgetting to use proper and consistent units throughout a pipeline, are often found when communication between ETL architects and the end-users (such as engineers and data analysts) is weak. Using standard formats for things like time-stamps (e.g., ISO 8601 [88]) is a common way to build interoperable data stores, but care must be taken to account for both present and future data storage needs of the enterprise when deciding to adopt formats or processes. One way to account for these data storage and transformation needs can be the use of standard data exchange protocols [89] and manufacturing information standards [90]. Protocols should be designed through strong relationships between management, planners, and data architects, along with the providers of any digital tools being used. Because experimentation with digital pipelines can be low-risk in the early stages (i.e. does not immediately impact production), it may be worthwhile to allow acceptable errors during technology adoption phases while exploring possible ETL architectures. This ensures that the final implemented design will correctly fit organizational needs, while exploiting the most recent techniques in ETL's rapidly shifting landscape.

Because planning for future storage and processing need is so critical, cloud computing through services like Amazon Web Services (AWS) [91] or Microsoft's Azure platform [92] present opportunities for flexible expansion of capabilities, that can scale with the needs of a system as it grows. Likewise, understanding exactly what functionality does and does not exist in adopted or purchased software stacks is key to planning ahead. If an engineer creates custom software to facilitate specific needs of an organization, good documentation practices are necessary for future understanding of the software. However, the engineer may never create this documentation due to time constraints, thus leading to the innate knowledge of that software retiring when they leave the company. By encouraging knowledge transfer

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

TABLE 4. Discover Maintenance Needs Tasks: Errors and Mitigations

Sub-task	Example-Error Description	Error Type		
Data Extraction, Transformation, and Loading (ETL) Tasks				
Extract	Inappropriate data quantity or type	SB		
	Not planning for volume	SB		
	Collect wrong data for desired analysis		RB	
Transform	Dimensions/units/feature compatibilty error	SB		
	Discard potentially useful data		RB	
Load	Choosing wrong hardware		RB	
	Misunderstanding software tools			KB
Modeling and Analytics Tasks				
Detect Trends	Not noticing trends	SB		
	Detecting false trends (overfitting)	SB		
	Insufficient communication of trends	SB		
Define Patterns	Inappropriate model or modeling assumptions		RB	
	Misinterpret correlation as causation		RB	
Identify Causality	Unknown relationships driving unknown failure modes			KB
	Lack framework for synthesizing model output into actionable strategies			KB

(e.g., to recent hires) and guideline creation, risk of relying on customized software and code for critical tasks can be reduced.

5.3 Modeling and Analytics Task Errors

Once data is loaded for analysis, it is analyzed through use of statistical summaries, model training, and data visualization. While there is no universal procedure for data analytics, there are practices to follow when modeling and analyzing data: 1) detection of trends in data, 2) defining of patterns between data types, and 3) identification of causal relationships and application potential. These practices map well with the GEMS, since the goals of each stage are similar to the goals of a problem solver — noticing a problem, noticing patterns that fix the problem, then understanding the mechanisms that cause the problem.

SB *Bearing sensor data from the past 5 months can be accessed via disparate spreadsheets, which indicate nominal system health over time, despite a 2-week period of increasing vibration* (Detect Trends Sub-task)

RB *Analyst estimates tool wear overhead with a physics-based model that calculates a mean time to failure metric for a CNC part; this model is not calibrated for one of the required depth-of-cut + diameter combinations* (Define Patterns Sub-task)

KB *A neural network trained with infrared maps of steel heat predicts a quality drop is imminent. The analyst is unable to determine a probable cause of the failure mode from the model, despite previously good model accuracy.* (Identify Causality Sub-task)

The first example is an oversight and can be addressed by making the data easier to see and accessible to more people. Numerous no cost, open source tools exist to perform initial statistical summary of data for reporting the key indicators of a need for work, and to visualize the data in ways that make trend detection at multiple time- and length-scales much more obvious. Starting to standardize or automate data pipelines that explicitly result in basic plots for machine performance summaries can be beneficial in building trust in an automated system [93, 94]. Addressing these errors is intertwined with proper ETL solutions, as discussed above: having a controlled data repository that structures and cross-links information makes designing and deploying such visualizations much easier. When well designed ETL is combined with dashboards (typically centralized displays of real-time streaming information), these types of well designed visualizations can take steps into mitigating RB and KB errors, by gathering disparate data sources into a single, easy-to-reference location which is accessible to multiple people. This makes the error of missed trends much less likely, and informs the creation of new rules for standard work.

A second major mitigation strategy, especially for the RB analysis errors, is the use of data-driven or hybrid data + physics models, for the detection and exploitation of patterns in observed

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian. "Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing." Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

equipment or system behavior for predicting health or performance. These are typically considered as part of Prognostics and Health Monitoring (PHM), a rapidly advancing sub-field of reliability engineering as described in Section 3.1. A key trade-off for using such models is that while high-accuracy predictions can be made when high quality and high quantity data is available, the models are not always interpretable, can be over-fitted, and may not be indicating causal (but rather coincidental) links between inputs and outputs. It is obviously better to be alerted to a prediction of failure than not, but if actionable strategies based on causal relationships are required, significantly more effort may be needed. False positive alarms reduce trust in the analysts and their predictions [95, 96]. Some forms of semantic and causal reasoning is possible, perhaps through design of custom ontologies or high-fidelity physics simulations, but implementing these tailor-made solutions presents a barrier, in infrastructure, labor, and research costs. Fortunately, expert knowledge can sometimes be applied to identify causation.

Based on this, investment only in the analysis stage starts out with high potential returns, but reaches a horizon as the needed technology to infer context and causality reaches the edge of the state-of-the-art. Readily-available technologies can assist analysts in addressing SB and RB errors is an efficient way to encourage them to use their own domain expertise in determining causality and potential strategies. This lays the groundwork to enable higher impact improvements in KB-intensive tasks, like execution and scheduling.

6 DISCUSSION

The previous sections provide a high level task and error analysis of the maintenance management workflow. Some common errors are classified according to Reason's GEMS framework (skill-, rule-, and knowledge-based errors). Careful consideration is made to distinguish between structured versus unstructured job tasks and errors. While the tasks and errors had much overlap, often they were performed by different roles within the organization and at different time scales. The errors are largely the same, but they occur more often with unstructured work. Unstructured jobs require decisions made in near real time by roles in the organization that are not meant to be making these decisions (e.g., a maintainer estimating severity of a failure on the fly) and in high stress situations (e.g., during a machine failure that can lead to production impacts). Shifting towards a more structured maintenance paradigm is important for an organization's success with new technology insertion. The steps provided in this paper are a beginning in this direction, however, as discussed, technological solutions are not the only mitigation strategy; cultural shifts are necessary as well [97].

Mitigation strategies are discussed for commonly occurring errors, independent of structured or unstructured jobs. These mitigation strategies range from necessary cultural changes to advanced AI solutions, however, these errors and mitigations do not represent every possible situation at different manufacturing facilities. How does a manufacturer repeat this same process and how should they implement their own technological solutions?

If one approaches modernizing a factory with new digital technological solutions as a problem solving situation, in a similar process to GEMS and the above discussion, the first step begins when stakeholders in the organization begin to perform an attentional check [25]. This step must happen before any problem can be solved because, by definition, this check identifies SB errors that are occurring without conscious recognition that something is wrong by the human actor.

For example, an operator not noticing an alarm that indicates failure is a potential SB error identified in Table 2. One solution to this problem, could be to install a sensor visualization dashboard to display the performance of the system for many to view. This solution could potentially solve the problem, but requires new sensors, logic for failure identification, visualization packages, etc. Does this solution always mitigate the error of not noticing a failure? By needing to capture new data and create a new visualization solution, how can one be sure the visualization is optimal to clearly indicate failure so the error does not occur? If the operator can miss an alarm, he/she can very easily not notice an icon in a visualization dashboard. It might not come to light that this solution is poor until a high investment in both time and cost is sunk into the project. A low technology solution might be a better answer to initially solve this error. For example, implementing a prototype visualization using existing data sources to ensure the operator can adapt to the new technology, or instilling a cultural enabler that could include a buddy system, so the operator can learn from more experienced colleagues could alleviate this error.

One of the most common causes for technology implementation failures, such as new CMMS, is an inability to make the necessary cultural changes [26]. Too often proponents of new software systems believe that the software implementation is the change, rather than putting effort in appropriate organizational change management processes to support the installation. For example, if operators and technicians are making many SB errors, such as not remembering significant symptoms from a maintenance job, a CMMS system will not immediately solve this problem. These types of errors often occur because the technician has no incentive to enter correct, long-form data about the maintenance job. In fact, the poor organizational culture encourages bad data entry, as these technicians are judged on how well and how quickly they solve the problem, not on the quality of the data [49]. If a CMMS system was installed, without addressing the SB errors, the technicians would still follow bad data entry practices, albeit on a much more expensive system. However, by discovering and attempting to mitigate SB errors, we enable a more efficient investigation into more emerging, sophisticated technologies that hold greater promise to automate systems and

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

more directly assist human decision making.

As SB errors are mitigated, a trust for new technologies builds. By alleviating the SB errors, the differences between the RB and KB errors will also emerge. It is typically the most knowledge intensive tasks for which humans are required, making these errors some of the most difficult to completely mitigate with new technologies. The next step after discovering and mitigating SB errors is investigating the emerging technologies for RB errors.

RB errors, by their nature, involve patterns and rules that are misapplied or an inappropriate rule. The avenues to identify the occurrence of rule-based errors include the use of digital pattern recognition and recall processes. For example, routine tasks can be aided through machine learning technologies that can learn the important features of the work. This technology augments the planner, scheduler, engineer and technician, who can use the knowledge to make appropriate decisions and focus on other tasks in their job.

Once the SB and RB errors are mitigated, manufacturers can attempt to address KB errors. As stated above, KB errors are difficult to mitigate with automation and are better suited for augmentation technologies that aid the human in the task. For example, imagine a technician attempts to solve a problem that he or she has never encountered before. To completely replace the human actor, in this scenario, with a robot is not realistic with the current technology solutions. It may be cost and time-effective to investigate AR solutions that can link with more experienced technicians and Computer Aided Design (CAD) drawings of the asset to visualize and talk through the current problem. However, while the solutions themselves might be low cost, creating an environment that connects CAD drawings, technicians, and visualization tools with assisted reality headsets, is often difficult for manufacturers to tackle if this is the first SM technology they employ.

While this procedure of error identification and technology mapping can help manufacturers, how can researchers push forward and create solutions that are used by manufacturers? Researchers are needed to reduce the cost of entry to these solutions, both in time, monetary cost, and required expertise. The exercise of identifying tasks and errors can leads to a better understanding of a manufacturer's trouble areas and provide more concrete use cases for researchers; however, scenarios are often not enough for some data-driven techniques. Realistic datasets, that are analogous to the data that occurs during maintenance, are necessary to train and prepare the data-driven models [98]. These types of datasets would support the development of open source data analysis and visualization tools that can greatly benefit manufacturers.

As the technologies are further developed and current technologies are deployed, guidelines for when and how to use various technologies are necessary. This paper ultimately provides guidance on what types of errors are dominant throughout the maintenance procedure, but stops short of discussing at length the pros and cons of each technology solution. Researchers can create and contribute to standard guidelines on what tools work and why for specific types of manufacturing datasets and problems. Guidance is also required to determine how to turn the outputs of the data analysis tools into actionable intelligence in a consistent manner. Lastly, manufacturers need to share their success stories in implementing these technologies for maintenance management. As shown in [2], the ROI of Smart Manufacturing technology implementation in maintenance ranges from 15 % to 98 %. As many manufacturers are nervous of the cost of these technologies, more rigorous studies of ROI are necessary to pave the way for other manufacturers. This paper can provide a first step in a Smart Manufacturing journey in maintenance.

7 CONCLUSIONS AND FUTURE WORK

This paper analyzes each step of the maintenance workflow: both reviewing current industrial implementations of research for each maintenance activity and providing a framework for determining the most cost effective points of entry for emerging technologies in Smart Manufacturing. The maintenance activities are broken down by tasks and potential errors are identified using Reason's taxonomy. The errors are classified according to Rasmussen's skill-, rule-, knowledge- performance model. This classification provides a framework to discuss the most effective areas to introduce emerging Smart Manufacturing technologies. Low- technology solutions, particularly cultural changes, can sometimes be employed to rectify skill-based errors; AI-driven solutions may solve rule-based errors; and knowledge-based errors will need high effort, high cost, and high fidelity system models to pull together many disparate data sources that form the human expertise.

Several potential areas for future exploration follow from this work:

More complete task analyses of the maintenance process — Further human reliability research should provide a more sophisticated breakdown of tasks, sub-tasks, and the associated potential errors. This type of analysis is necessary to fully recognize the relationship between the human actors in maintenance and the technology solutions applicable to a manufacturing facility. Using a more complete task analysis on the maintenance workflow, will allow researchers to better understand the interrelationships of the human and technology within the maintenance workflow.

Systematic error identification and tracking — Solutions are needed to provide manufacturers with guidance on how to perform this analysis across the entire manufacturing facility. A key aspect of recognizing error severity is the determination of key performance shaping factors: environmental or other con-

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

textual influences that modify error likelihood (also called common performance conditions, see [29]). Having a repository of manufacturing maintenance errors, perhaps taking a cue from the U.S. Nuclear Regulatory Commission Human Event Repository and Analysis (HERA) database [99], could prove useful for more efficient error modeling, going forward.

Human models and assistance through machine learning — In areas like maintenance that require human engagement, and tend to generate smaller data compared to other domains, up-and-coming advances in machine learning that can handle a lack of large training datasets will have an understated impact on our ability to model and assist relevant aspects of human behavior. These types of models, whether focused on reliability prediction, ergonomic optimization, or performance measures, are becoming possible through hybridized learning techniques, which exploit existing basic knowledge about some model while still adapting to new circumstances in reasonable ways. This provides a mechanism for ML to assist less experienced practitioners in learning their domain: "intelligence augmentation" over "artificial intelligence" [100]. Techniques like restricting predictions to a learned space of useful results [101, 102], discovering computational models for difficult-to-quantify user preference in decision making [103, 104], and many more, can be directly applied to better model and assist maintenance practitioner's diagnostic and execution behavior.

Guidelines on tools that are available in Smart Manufacturing and the potential benefits and drawbacks of each method or tool — This paper provided examples of tools that are available in industry, but did not enumerate every potential Smart Manufacturing technology. More work is necessary to discuss how and when to use specific techniques for the appropriate problem in manufacturing, including not only potential benefits but also drawbacks.

Reference datasets from manufacturers for analysis comparison — Within manufacturing, publicly available datasets mimicking real world scenarios are lacking. Without these datasets, it is difficult for analysis to provide solutions that works in real manufacturing environments.

Standard guidelines on how to perform this analysis consistently within manufacturing — While this paper provides the first steps in this process, this work can be forwarded through standard organizations to provide simple-to-follow guidance allowing manufacturers to perform this analysis on their own in a structured way.

As the factories of the future become more and more automated, the skills required to support manufacturing operations will shift from operations to maintenance. In this environment, manufacturers need to understand the best place to start with implementing these emerging technologies. The optimal path forward with technology for maintenance is not to replace a human in the workflow. A solution that augments the human's abilities will take advantage of the human's cognitive capability while removing the reducing errors. In fact, in the future, Intelligence Augmentation (IA) might become a more practical approach, compared to AI, since it supplements human's cognitive process at different levels of Bloom's Taxonomy while keeping the human at the center of the decision-making process [105]. This paper allows manufacturers to stop asking how to get "smart", but instead allows manufacturers to ask how can we "smartly" implement new technologies in maintenance with the highest probability for success by accounting for the errors the technology will alleviate.

DISCLAIMER

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

ACKNOWLEDGMENT

Contributions from the University of Western Australia for this work was supported in part by the BHP Fellowship for Engineering for Remote Operations. The authors would like to thank Moneer Helu, Qing Chang, and Vijay Srinivasan for insightful feedback during the preparation of this manuscript.

REFERENCES

[1] IEC, 2011. *AS IEC 60300.3.11 Dependability management Application guide - Reliability-centred Maintenance*. Geneva Switzerland.

[2] Thomas, D. S., 2018. The costs and benefits of advanced maintenance in manufacturing. Tech. rep.

[3] Feldman, K., Sandborn, P., and Jazouli, T., 2008. "The analysis of return on investment for phm applied to electronic systems". In Prognostics and Health Management, 2008. PHM 2008. International Conference on, IEEE, pp. 1–9.

[4] Drummond, C., and Yang, C., 2008. "Reverse engineering costs: How much will a prognostic algorithm save". In International Conference on Prognostics and Health Management, Denver, CO.

[5] Yang, C., and Létourneau, S., 2007. "Model evaluation for prognostics: Estimating cost saving for the end users". In Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on, IEEE, pp. 304–309.

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

[6] Nowlan, F. S., and Heap, H. F., 1978. Reliability-centered maintenance. Tech. rep., United Air Lines Inc San Francisco Ca.
[7] IEC, 2016. *AS IEC 60300.3.14 Dependability management Application guide - Maintenance and maintenance support*. Geneva Switzerland.
[8] Kelly, A., 2006. *Strategic maintenance planning*, Vol. 1. Elsevier, Oxford.
[9] Kelly, A., 1997. *Maintenance organization and systems*. Butterworth-Heinemann.
[10] Palmer, D., 1999. *Maintenance planning and scheduling handbook*. McGraw-Hill Professional Publishing.
[11] GFMAM, 2016. *Maintenance Framework*. London, England.
[12] SMRP, 2009. *SMRP Best Practice - Maintenance —& Reliability Body of Knowledge*, 5th. ed. Society of Maintenance and Reliability Professionals, Atlanta, GA.
[13] Nakajima, S., 1988. "Introduction to tpm: Total productive maintenance (preventative maintenance series)". *Hardcover. ISBN 0-91529-923-2*.
[14] Blanchard, B. S., 1997. "An enhanced approach for implementing total productive maintenance in the manufacturing environment". *Journal of quality in Maintenance Engineering,* ***3***(2), pp. 69–80.
[15] McKone, K. E., Schroeder, R. G., and Cua, K. O., 2001. "The impact of total productive maintenance practices on manufacturing performance". *Journal of operations management,* ***19***(1), pp. 39–58.
[16] Smith, R., and Hawkins, B., 2004. *Lean maintenance: reduce costs, improve quality, and increase market share*. Elsevier.
[17] Mostafa, S., Dumrak, J., and Soltan, H., 2015. "Lean maintenance roadmap". *Procedia Manufacturing,* ***2***, pp. 434–444.
[18] Jin, Xiaoning, Siegel, David, Weiss, Brian A., Gamel, Ellen, Wang, Wei, Lee, Jay, and Ni, Jun, 2016. "The present status and future growth of maintenance in us manufacturing: results from a pilot survey". *Manufacturing Rev.,* ***3***, p. 10.
[19] Alsyouf, I., 2007. "The role of maintenance in improving companies' productivity and profitability". *International Journal of production economics,* ***105***(1), pp. 70–78.
[20] Mobley, R. K., 2002. *An introduction to predictive maintenance*. Elsevier.
[21] Vogl, G. W., Weiss, B. A., and Helu, M., 2016. "A review of diagnostic and prognostic capabilities and best practices for manufacturing". *Journal of Intelligent Manufacturing*, pp. 1–17.
[22] Director, Mission Assurance, and Director of Human Performance, Training Biosystems, 2011. FY 011 Department of Defense human systems integration management plan. Tech. rep.
[23] O'HARA, J. M., and Brown, W., 2004. Incorporation of human factors engineering analyses and tools into the design process for digital control room upgrades. Tech. rep., BROOKHAVEN NATIONAL LABORATORY (US).
[24] Reason, J., 1990. *Human error*. Cambridge university press.
[25] Rasmussen, J., 1983. "Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models". *IEEE transactions on systems, man, and cybernetics*(3), pp. 257–266.
[26] Thomas, S. J., 2005. *Improving maintenance and reliability through cultural change*. Industrial Press Inc.
[27] Kirwan, B., Gibson, H., Kennedy, R., Edmunds, J., Cooksley, G., and Umbers, I., 2004. "Nuclear action reliability assessment (nara): a data-based hra tool". In Probabilistic safety assessment and management, Springer, pp. 1206–1211.
[28] Gertman, D., Blackman, H., Marble, J., Byers, J., Smith, C., et al., 2005. "The spar-h human reliability analysis method". *US Nuclear Regulatory Commission*.
[29] Hollnagel, E., 1998. *Cognitive reliability and error analysis method (CREAM)*. Elsevier.
[30] Dekker, R., 1996. "Applications of maintenance optimization models: a review and analysis". *Reliability engineering & system safety,* ***51***(3), pp. 229–240.
[31] Dekker, R., and Scarf, P. A., 1998. "On the impact of optimisation models in maintenance decision making: the state of the art". *Reliability Engineering & System Safety,* ***60***(2), pp. 111–119.
[32] Márquez, A. C., 2007. *The maintenance management framework: models and methods for complex systems maintenance*. Springer Science & Business Media.
[33] Jardine, A. K., and Tsang, A. H., 2013. *Maintenance, replacement, and reliability: theory and applications*. CRC press.
[34] Alrabghi, A., and Tiwari, A., 2015. "State of the art in simulation-based optimisation for maintenance systems". *Computers & Industrial Engineering,* ***82***, pp. 167–182.
[35] Ribeiro, M., Silveira, J., and Qassim, R., 2007. "Joint optimisation of maintenance and buffer size in a manufacturing system". *European Journal of Operational Research,* ***176***(1), pp. 405–413.
[36] Chang, Q., Ni, J., Bandyopadhyay, P., Biller, S., and Xiao, G., 2007. "Maintenance opportunity planning system". *Journal of Manufacturing Science and Engineering,* ***129***(3), pp. 661–668.
[37] Li, Y., Tang, Q., Chang, Q., and Brundage, M. P., 2017. "An event-based analysis of condition-based maintenance decision-making in multistage production systems". *International Journal of Production Research,* ***55***(16), pp. 4753–4764.
[38] Hoffman, M., Song, E., Brundage, M., and Kumara, S.,

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

2018. "Condition-based maintenance policy optimization using genetic algorithms and gaussian markov improvement algorithm". In PHM Society Conference, Vol. 10.

[39] Kenné, J. P., and Nkeungoue, L., 2008. "Simultaneous control of production, preventive and corrective maintenance rates of a failure-prone manufacturing system". *Applied numerical mathematics,* ***58***(2), pp. 180–194.

[40] Song, D.-P., 2009. "Production and preventive maintenance control in a stochastic manufacturing system". *International Journal of Production Economics,* ***119***(1), pp. 101–111.

[41] de Castro, H. F., and Cavalca, K. L., 2006. "Maintenance resources optimization applied to a manufacturing system". *Reliability Engineering & System Safety,* ***91***(4), pp. 413–420.

[42] Lee, J., Lapira, E., Bagheri, B., and Kao, H.-a., 2013. "Recent advances and trends in predictive manufacturing systems in big data environment". *Manufacturing Letters,* ***1***(1), pp. 38–41.

[43] Bajaj, M., and Hedberg Jr, T., 2018. "System lifecycle handler—spinning a digital thread for manufacturing". In INCOSE International Symposium, Vol. 28, Wiley Online Library, pp. 1636–1650.

[44] Rosen, R., Von Wichert, G., Lo, G., and Bettenhausen, K. D., 2015. "About the importance of autonomy and digital twins for the future of manufacturing". *IFAC-PapersOnLine,* ***48***(3), pp. 567–572.

[45] Monostori, L., Váncza, J., and Kumara, S. R., 2006. "Agent-based systems for manufacturing". *CIRP Annals-Manufacturing Technology,* ***55***(2), pp. 697–720.

[46] Shen, W., Hao, Q., Yoon, H. J., and Norrie, D. H., 2006. "Applications of agent-based systems in intelligent manufacturing: An updated review". *Advanced engineering INFORMATICS,* ***20***(4), pp. 415–431.

[47] Hu, X., Griffin, M., Yeo, G., Kanse, L., Hodkiewicz, M., and Parkes, K., 2018. "A new look at compliance with work procedures: An engagement perspective". *Safety science,* ***105***, pp. 46–54.

[48] Molina, R., Unsworth, K., Hodkiewicz, M., and Adriasola, E., 2013. "Are managerial pressure, technological control and intrinsic motivation effective in improving data quality?". *Reliability Engineering System Safety,* ***119***, pp. 26 – 34.

[49] Unsworth, K., Adriasola, E., Johnston-Billings, A., Dmitrieva, A., and Hodkiewicz, M., 2011. "Goal hierarchy: Improving asset data quality by improving motivation". *Reliability Engineering System Safety,* ***96***(11), pp. 1474 – 1481.

[50] Singh, S., Kumar, R., and Kumar, U., 2015. "Applying human factor analysis tools to a railway brake and wheel maintenance facility". *Journal of Quality in Maintenance Engineering,* ***21***(1), pp. 89–99.

[51] Reason, J., and Hobbs, A., 2017. *Managing maintenance error: a practical guide*. CRC Press.

[52] Kanse, L., Parkes, K., Hodkiewicz, M., Hu, X., and Griffin, M., 2018. "Are you sure you want me to follow this? a study of procedure management, user perceptions and compliance behaviour". *Safety science,* ***101***, pp. 19–32.

[53] Morkos, B., Taiber, J., Summers, J., Mears, L., Fadel, G., and Rilka, T., 2012. "Mobile devices within manufacturing environments: a bmw applicability study". *International Journal on Interactive Design and Manufacturing (IJIDeM),* ***6***(2), pp. 101–111.

[54] Aerospace, T., 2018. *"Fountx"*. Available at http://fountx.com/about-fountx/. Accessed 10-27-18.

[55] HoloGroup, 2018. *"HoloLens"*. Available at https://holo.group/en/corporate/. Accessed 10-27-18.

[56] Google, 2018. *"Google Glass"*. Available at https://x.company/glass/. Accessed 10-27-18.

[57] Smoker, T. M., French, T., Liu, W., and Hodkiewicz, M. R., 2017. "Applying cognitive computing to maintainer-collected data". In System Reliability and Safety (ICSRS), 2017 2nd International Conference on, IEEE, pp. 543–551.

[58] Sexton, T., Brundage, M. P., Morris, K., and Hoffman, M., 2017. "Hybrid datafication of maintenance logs from ai-assisted human tags". IEEE Big Data 2017, pp. 1–8.

[59] Brundage, M. P., Sexton, T., Moccozet, S., Hoffman, M., and Morris, K., 2018. "Developing maintenance key performance indicators from maintenance work order data". In Proceedings of the ASME 2018 International Manufacturing Science and Engineering Conference, MSEC2018, American Society of Mechanical Engineers.

[60] Sexton, T., Hodkiewicz, M., Brundage, M. P., and Smoker, T., 2018. "Benchmarking for keyword extraction methodologies in maintenance work orders". In PHM Society Conference, Vol. 10.

[61] Ebrahimipour, V., and Yacout, S., 2015. "Ontology-based schema to support maintenance knowledge representation with a case study of a pneumatic valve". *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* ***45***(4), pp. 702–712.

[62] Mazzola, L., Kapahnke, P., Vujic, M., and Klusch, M., 2016. "Cdm-core: A manufacturing domain ontology in owl2 for production and maintenance". In Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2016, SCITEPRESS - Science and Technology Publications, Lda, pp. 136–143.

[63] Wallace, E., Kiritsis, D., Smith, B., Will, C., et al., 2018. "The industrial ontologies foundry proof-of-concept project". In IFIP International Conference on Advances in Production Management Systems, Springer, pp. 402–409.

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

[64] IOF, 2018. *"IOF"*. Available at https://sites.google.com/view/industrialontologies/home. Accessed 10-10-18.

[65] Roberts, K. H., 1990. "Managing high reliability organizations". *California management review,* ***32***(4), pp. 101–113.

[66] Kawamoto, K., Houlihan, C. A., Balas, E. A., and Lobach, D. F., 2005. "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success". *Bmj,* ***330***(7494), p. 765.

[67] ISO, 2006. *ISO 60712 Analysis techniques for system reliability - procedure for failure mode and effects analysis*. Geneva, Switzerland.

[68] Rausand, Martin, and Vatn, Jorn, 2008. "Reliability centred maintenance". In *Complex system maintenance handbook*. Springer, pp. 79–108.

[69] Moubray, J., 2007. *Reliability-centered Maintenance RCM II*, 2nd. ed. Butterworth-Heinemann, Oxford.

[70] SAE, 2011. *SAE JA1012 A guide to the Reliability-centered maintenance (RCM) Standard*. London.

[71] Bertling, L., 2002. "Reliability-centred maintenance for electric power distribution systems". PhD thesis, Elektrotekniska system.

[72] Schlabbach, R., and Berka, T., 2001. "Reliability-centred maintenance of mv circuit-breakers". In Power Tech Proceedings, 2001 IEEE Porto, Vol. 4, IEEE, pp. 5–pp.

[73] Mokashi, A., Wang, J., and Vermar, A., 2002. "A study of reliability-centred maintenance in maritime operations". *Marine Policy,* ***26***(5), pp. 325–335.

[74] Igba, J., Alemzadeh, K., Anyanwu-Ebo, I., Gibbons, P., and Friis, J., 2013. "A systems approach towards reliability-centred maintenance (rcm) of wind turbines". *Procedia Computer Science,* ***16***, pp. 814–823.

[75] Tu, P. Y., Yam, R., Tse, P., and Sun, A., 2001. "An integrated maintenance management system for an advanced manufacturing company". *The International Journal of Advanced Manufacturing Technology,* ***17***(9), pp. 692–703.

[76] Jonsson, P., 1997. "The status of maintenance management in swedish manufacturing firms". *Journal of Quality in Maintenance Engineering,* ***3***(4), pp. 233–258.

[77] Astfalck, L., Hodkiewicz, M., Keating, A., Cripps, E., and Pecht, M., 2016. "A modelling ecosystem for prognostics". In Proceedings of the 2016 annual conference of the Prognostics and Health Management Society, PHM Society.

[78] Sikorska, J., Hodkiewicz, M., and Ma, L., 2011. "Prognostic modelling options for remaining useful life estimation by industry". *Mechanical Systems and Signal Processing,* ***25***(5), pp. 1803–1836.

[79] Sankararaman, S., Saxena, A., and Goebel, K., 2014. "Are current prognostic performance evaluation practices sufficient and meaningful?". In Proceedings of the 2014 annual conference of the Prognostics and Health Management Society, PHM Society.

[80] Kwon, D., Hodkiewicz, M. R., Fan, J., Shibutani, T., and Pecht, M. G., 2016. "Iot-based prognostics and systems health management for industrial applications". *IEEE Access,* ***4***, pp. 3659–3670.

[81] Dassault, 2018. *"Dassault Systemes"*. Available at https://www.3ds.com/. Accessed 10-10-18.

[82] GE, 2018. *"GE Predix"*. Available at https://www.ge.com/digital/iiot-platform. Accessed 10-10-18.

[83] PTC, 2018. *"PTC Thingworx"*. Available at https://www.ptc.com/en/products/iot/thingworx-platform. Accessed 10-10-18.

[84] Automation, I., 2018. *"Ignition"*. Available at https://inductiveautomation.com. Accessed 10-10-18.

[85] Siemens, 2018. *"Siemens Mindsphere"*. Available at https://www.ptc.com/en/products/iot/thingworx-platform. Accessed 10-10-18.

[86] Lee, J., Ni, J., Djurdjanovic, D., Qiu, H., and Liao, H., 2006. "Intelligent prognostics tools and e-maintenance". *Computers in industry,* ***57***(6), pp. 476–489.

[87] Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., and Siegel, D., 2014. "Prognostics and health management design for rotary machinery systems—reviews, methodology and applications". *Mechanical systems and signal processing,* ***42***(1-2), pp. 314–334.

[88] Klyne, G., and Newman, C., 2002. Date and time on the internet: Timestamps. Tech. rep.

[89] Foundation, O., 2018. *"OPC UA"*. Available at https://opcfoundation.org/. Accessed 10-27-18.

[90] Institute, M., 2018. *"MTConnect"*. Available at https://www.mtconnect.org/. Accessed 10-27-18.

[91] Amazon, 2018. *"Amazon Web Services"*. Available at https://aws.amazon.com/. Accessed 10-27-18.

[92] Microsoft, 2018. *"Microsoft Azure"*. Available at https://azure.microsoft.com/en-us/. Accessed 10-27-18.

[93] Jo, J., Huh, J., Park, J., Kim, B., and Seo, J., 2014. "Livegantt: Interactively visualizing a large manufacturing schedule". *IEEE transactions on visualization and computer graphics,* ***20***(12), pp. 2329–2338.

[94] Xu, P., Mei, H., Ren, L., and Chen, W., 2017. "Vidx: Visual diagnostics of assembly line performance in smart factories". *IEEE transactions on visualization and computer graphics,* ***23***(1), pp. 291–300.

[95] Phillips, J., Cripps, E., Lau, J. W., and Hodkiewicz, M., 2015. "Classifying machinery condition using oil samples and binary logistic regression". *Mechanical Systems and Signal Processing,* ***60***, pp. 316–325.

[96] Bliss, J. P., and Dunn, M. C., 2000. "Behavioural impli-

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian.
"Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing."
Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

cations of alarm mistrust as a function of task workload". *Ergonomics,* ***43***(9), pp. 1283–1300.

[97] Chor, K. H. B., Wisdom, J. P., Olin, S.-C. S., Hoagwood, K. E., and Horwitz, S. M., 2015. "Measures for predictors of innovation adoption". *Administration and Policy in Mental Health and Mental Health Services Research,* ***42***(5), Sep, pp. 545–573.

[98] Sikorska, J., Hodkiewicz, M., De Cruz, A., Astfalck, L., and Keating, A., 2016. "A collaborative data library for testing prognostic models". In 3rd Eur. Conf. Prognostics Health Manage. Soc.

[99] Hallbert, B., Boring, R., Gertman, D., Dudenhoeffer, D., Whaley, A., Marble, J., Joe, J., and Lois, E., 2006. "Human event repository and analysis (hera) system, overview". *US Nuclear Regulatory Commission, Washington DC, Tech. Rep. NUREG/CR-6903*.

[100] Skagestad, P., 1993. "Thinking with machines: Intelligence augmentation, evolutionary epistemology, and semiotic". *Journal of Social and Evolutionary Systems,* ***16***(2), pp. 157–180.

[101] Chen, W., Fuge, M., and Chazan, J., 2017. "Design manifolds capture the intrinsic complexity and dimension of design spaces". *Journal of Mechanical Design,* ***139***(5), p. 051102.

[102] Duvenaud, D., 2014. "Automatic model construction with gaussian processes". PhD thesis, University of Cambridge.

[103] Sexton, T., and Ren, M. Y., 2017. "Learning an optimization algorithm through human design iterations". *Journal of Mechanical Design,* ***139***(10), p. 101404.

[104] Gonzalez, J., Dai, Z., Damianou, A., and Lawrence, N. D., 2016. "Bayesian optimisation with pairwise preferential returns". In NIPS Workshop on Bayesian Optimization.

[105] BLOOM'S, T. M. E., 1965. *Bloom's taxonomy of educational objectives*. Longman.

Brundage, Michael; Sexton, Thurston; Hodkiewicz, Melinda; Morris, Katherine; Arinez, Jorge; Ameri, Farhad; Ni, Jun; Xiao, Guoxian. "Where Do We Start? Guidance for Technology Implementation in Maintenance Management for Manufacturing." Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference, MSEC 2019 MSEC2019, Erie, PA, United States. June 10, 2019 - June 14, 2019.

Selecting Optimal Data for Creating Informed Maintenance Decisions in a Manufacturing Environment

Don't Drown in Trash: Curating 'Minimum Viable' Data Sets

Authors:
Michael Sharp, Michael P. Brundage, Timothy Sprock, Brian A. Weiss

Abstract

Data availability within a manufacturing enterprise directly drives the ability of decision makers to effectively function and operate. The information needs of decision makers can vary greatly, based not only on the level at which the decision is being made, but also the perspective and desired effect of that decision. For example, an equipment-level operator needs direct knowledge of that equipment's condition when deciding whether to operate that machine today; a production manager needs to know the number of operational machines when planning system-level operations; a maintenance manager needs knowledge of what maintenance tasks are in the queue and the availability of technicians. Although each decision is related, the information required to support each decision is distinct, and generated from sources that are often independent of one another. The granularity of information needed to make a decision is informed directly by what that decision is and any consequences of that decision. This paper discusses information and data requirements for maintenance decisions in manufacturing from multiple perspectives, including system, equipment, and component -level decisions. These decisions include both structured maintenance (planned and scheduled in advance of failures) and unstructured maintenance (performed immediately after a failure) decisions. The goal of this paper is to guide manufacturers who have limited resources to invest into a monitoring program to select a minimum viable set of data items to collect that support the decisions they want to make.

Introduction

The competitive edge in many industries, including manufacturing, is built upon fast, informed, and experienced decision making. Knowledge needed for informed decision-making differs based on the perspective and role of the decision maker. Decision support information requirements vary based on both the type of decisions being made and the level the decision affects, such as asset level, plant-wide level, or enterprise level. Reliability, maintenance, and operations planning all require specialized information resources to develop highly informed decisions in manufacturing facilities. What all of these levels and perspectives of decision making have in common is the need for information, i.e., data. However, not all facilities are equipped or have resources to devote to developing fully integrated or exhaustive data collection systems.

Sharp, Michael; Brundage, Michael; Sprock, Timothy; Weiss, Brian.
"Selecting Optimal Data for Creating Informed Maintenance Decisions in a Manufacturing Environment - Don't Drown in Trash: Curating 'Minimum Viable' Data Sets."
Paper presented at Model-Based Enterprise Summit 2019, Gaithersburg, MD, United States. April 1, 2019 - April 4, 2019.

Sources of data are varied in modern manufacturing facilities. Each day a wealth of potential information is generated from equipment, sensors, and routine activities performed by plant personnel. Unfortunately, even in facilities with existing data collection systems, a majority of this information is not being properly captured and documented in a manner that allows for proper utilization of that data. While ideally data would always be managed with end uses and goals in mind, that is often not feasible as new uses and needs evolve with changing technologies and environments of factory floors. To maximize resource utilization and effectiveness it is important to continually assess the minimal set of decision making information needs and verify that the manufacturing facility's current information capturing policies and technologies can support them. When considering areas and systems with 'critical information', the goal is to develop a plan to collect a minimally viable amount of data that allows sufficient characterization and modeling of the system for intelligent decision making without devoting resources to unneeded or ineffective information. Too much data collection adds unnecessary processing costs and time as well as increasing the demands on properly storing and curating the information. Too little data collection could increase uncertainty or erroneous assumptions leading to suboptimal planning and lost productivity. These inefficient and ineffective scenarios can be avoided by structuring monitoring and analysis activities to collect and curate the smallest amount of data that sufficiently answers decision support questions for maintenance and operations management.

This paper is the first in a series making recommendations to help manufacturers develop viable monitoring programs and reference data sets to support informed maintenance decision making at various levels of the enterprise. Before focusing on technology solutions to capture and store data, it is also important to understand what constitutes useful amounts and types of data. This requirement list is largely informed by the intended use of the data being gathered. Both qualitative and quantitative sources of information are needed to produce comprehensive assessments of the condition, capability, and capacity of systems at all levels of the manufacturing enterprise. Finding the minimum viable information set requires identifying how much and what kinds of data are needed to enable efficient informed decision making necessary for competitive operations.

Background and Motivation

Gathering and assessing information as it moves in a manufacturing facility is a never-ending task that aids in all levels of facility functions. In many instances, manufacturers benefit from both quantitative data coming from embedded/OEM (original equipment manufacturer) and third party sensors, as well as qualitative data coming from human operators. Tools for collecting, storing, and processing these sources of information have evolved to allow the generated data to contain both more content and volume [Lee et al 2018]. Technologies for visualizing and interpreting data have also grown, promoting further utilization, understanding, and justification of any decisions made within or about a facility [Sackett et al 2006]. However, a number of concerns still exist in practical implementations of decisions using these data sources including: 1) potential for data capture to impede normal operating procedures, 2) lack of data interoperability, and 3) lack of validation guidance.

Sharp, Michael; Brundage, Michael; Sprock, Timothy; Weiss, Brian.
"Selecting Optimal Data for Creating Informed Maintenance Decisions in a Manufacturing Environment - Don't Drown in Trash: Curating 'Minimum Viable' Data Sets."
Paper presented at Model-Based Enterprise Summit 2019, Gaithersburg, MD, United States. April 1, 2019 - April 4, 2019.

One concern when implementing information collection procedures and technologies is to ensure the information collection does not significantly interfere or impede normal operating procedures. This is true with both digital sensing, such as Industrial Internet of Things (IIoT) where computing and communication resources (bandwidth) for controlling the system may compete with data collection and transmission. Similarly, personnel performing value-add activities, such as completing maintenance tasks, often competes with time spent on filling out work orders and maintenance logs. While such procedures can provide large benefits for analysis and planning, they compete for time that could be spent performing other tasks.

An ideal reference data set would have balanced amounts of information representing all possible expected states of the equipment and facility. In most situations, creating this is impractical while maintaining normal operations. For example, many predictive diagnostic models at the asset or component level require one or more observed failures of a given type to characterize incipient fault symptoms. Generally, in real-world environments, every precaution would be made to prevent such failures making them impossible to observe and record. There is a need to develop different points/sources of data or model types to overcome this. Additionally, some system conditions could be exceedingly rare, or just not practical to enact at the time that the model is initially being created or tested. For these reasons, it is important to have mechanisms for adapting, updating, or replacing any information and decision support models as new data becomes available during the life of the system.

Unfortunately, even with modern data capturing technologies and procedures, it is rare for manufacturers to capture or have access to the 'perfect' data set for any given end goal. What limits the usefulness of a data set is not only the lack of balanced coverage of the information, but also a lack of annotation and context within data that is available. Ideally, when disparate sources of data relate to similar (or identical) kinds of system(s) or asset(s), those sources of data would be semantically linked or annotated in a manner that allows alignment and concurrent use for models and analysis. Rarely can a single source of information fully characterize a system or asset. Both quantitative and qualitative sources, such as maintenance reports [Sexton et al 2018] and sensors [Kong et al 2017] are needed for a full picture of the production facility. Manufacturers rely on standards for product information (e.g., STEP [ISO 10303], G-code [Kramer 2000], QIF [QIF 3.0, 2018]); equipment data (e.g., MTConnect [MTConnect Standard 2018]); and non-standardized data collected from Computerized Maintenance Management Systems (CMMS), as well as raw and processed sensor logs. Despite the availability of these data standards, the lack of interoperability of the associated data sources and integrated third party tools highlights practical concerns of linking dissimilar file formats. This also points to a need for intermediate platforms for information extraction and collection that can be accomplished in a practical way useful for various levels of informed decision making. More basic than information extraction, there is a need for a standardized manner to discover and/or assign connections between the files that can facilitate information discovery.

There are few standardized methods for determining how much and what kinds of data are necessary to build models or perform validation on decision support platforms at a given level of the enterprise. Even when restricting to a single decision level, such as the equipment or component level, available models each have their own requirements regarding active and historic data [Si 2011]. Adding to this, even if a

'perfect' data set were to exist that could translate information across all levels of the enterprise, no consistent guidance is given on how to turn this data into actionable intelligence for decision making.

When developing an information and decision support structure, there are two philosophies for approaching this: 'what is the minimum data I need to answer my questions?' versus 'what questions can I answer with the data I have available?'. The authors approach the issue from the first perspective, which implicitly requires decision makers to understand and focus on fewer, but higher impact questions. This builds from the idea that it is not realistic, useful, nor feasible to capture all information. Our research goal is a framework guiding the development of minimum viable data collection and storage that satisfies all critical decision support needs without over burdening employees or other resources with unnecessary tasks of curating and dissecting information of limited value. The next section examines some practical steps that can be used in developing both historical and ongoing data sets for characterization and modeling of system states for informed decision support.

Methodology

This paper explores methods for developing 'minimally viable' data sets required for informed, intelligent decision making. The first step in determining the minimally viable amount of data collection focuses on decomposing the functions and assets of the factory into linked levels and subdivisions that represent the physical and functional structure of the facility [Li et al 2018]. This architecture can be used to help model and identify the most critical links that either contain or transmit data and information needed to make useful observations about the state of the facility [Sharp 2018]. The ISA-88 and ISA-95 family of standards prescribe an enterprise hierarchy: field device (sensors and actuators), control device (control devices, controllers, embedded controllers), and station (machines, robots, intelligent logistics/material handling). For the purposes of this paper, the authors focus on a simplified hierarchy: system, equipment, and component levels. Some additional context and characterization of the component, equipment, and system is provided below.

Within the context of maintenance decisions, the "bottom" of the decomposition hierarchy is defined by the lowest repairable unit (LRU) and any associated performance or condition indicators that are monitored. The specific list and level of LRU assets will be unique to each enterprise, but the definition generally centers on the lowest level component that can be maintained, fixed, or replaced on site, and whose failure would have a negative impact on the site's performance or efficiency. Some examples of LRU could be bearings, sealed motors, hydraulic actuators, or other assets that are generally repaired or replaced on site. Given this list, it is important to note that although most LRUs are found at the component level, in certain situations this could be found at either the equipment or even system level. For example, if a specialized milling machine must be maintained by an outside contractor, it may only be necessary for the factory to monitor if the milling machine is maintaining high level performance indicators. A system level LRU could be a digital software system, such as a third party off the shelf CMMS that might simply be replaced with another if it fails to meet the facility's needs or proves faulty / obsolete.

Components are physical entities defined by a single, static functional capability controlled parametrically by well-defined inputs and expected effects. Each can be maintained (or replaced) independently of the equipment it's a part of. Components can be composed of other components (i.e., sub-components). Most LRU assets are found at the component level. Examples of components could be pumps, bearings, actuators, wiring, etc.

Equipment are composed of components and/or other equipment, and are described as "functionally complete" units. Information and data from them supports decision-making focused on real-time process execution. Concerns regarding the equipment level are embodied in supervisory control systems (SCADA), advanced-process control and optimization (APC-O) and programmable logic controllers (PLC). From the maintenance perspective, the availability and capability of the equipment is directly impacted by its components. While an equipment asset may be "maintained" or inspected, maintenance activities most often address the lower level components that comprise the equipment. Examples of equipment level assets could be milling machines, robotic manipulators, casting machines, or other assets that perform one or more tasks facilitating production and have components within them that can be replaced or repaired on site.

Systems are composed of equipment and subsystems and focus on completing one or more production tasks. Information relating to systems is used for decision-making focused on material/job flow, resource utilization and contention, and enterprise concerns such as throughput, cycle-time, quality, and cost. These concerns are often embodied in manufacturing operations management (MOM) and manufacturing execution systems (MES). Additionally from the maintenance perspective, systems are not individual units that can be directly maintained, but rather has its capability, capacity, and performance defined by its constituent units (subsystems and equipment assets). Work cells, work stations, production lines, or even full facilities could be considered systems and collections of subsystems from an enterprise level.

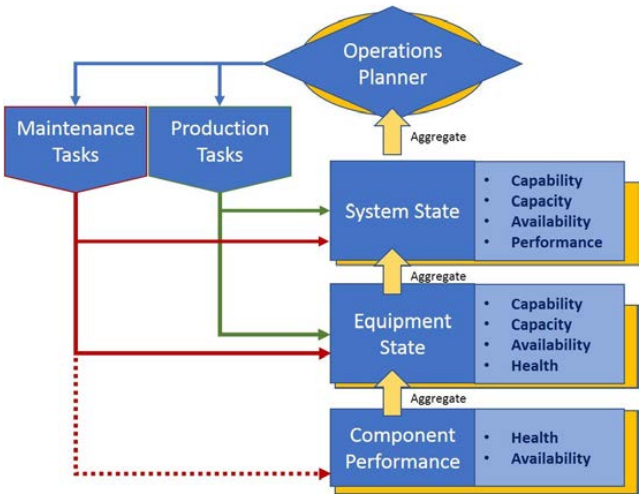


Figure 1: Information Level Flow Diagram

Each level of this relatively simple hierarchy (component, equipment, system) provides state, capability, or performance information that is aggregated upwards to inform the next level (Figure 1). In this simplified scheme, information from component and LRU level assets is propagated upwards to inform about the operations state and condition of the associated equipment. Equipment state and condition, in turn, is used to inform its system's operations and performance. The system state, aggregating many pieces of information, is input into an operations planner that prioritizes and schedules both production and maintenance tasks that are fed back through the various levels. Should additional information be required at a higher level than it is typically transmitted to, it would ideally be a simple matter to drill down and query any information set without losing any information. Figure 1 is a simplified flow chart that represents an implementation of this process. Alternatives or variants of this diagram could also exist, but the basic structure of condensing and feeding information upwards to a planner that then directs operations is the idea explored in this paper.

Figure 1 indicates that it is the primary job of the component level to populate maintenance tasks based on both condition- and calendar-based events for any given LRU. The equipment level information may also provide maintenance tasks that cross-cut multiple components, or may not have been easily discoverable based upon component level monitoring. Evaluating the criticality of each maintenance task is mostly performed with contextual information from the equipment level. Equipment level information also supports coordinating maintenance tasks that should be performed together. Coordinating maintenance tasks is possible when a planned operation provides an opportunity to performance additional maintenance while minimally impact operations, e.g., replacing a faulty valve when a machine is on a planned service outage for routine lubrication change. This coordination function can similarly be performed at the system level when managing linked equipment. Full maintenance scheduling is best accomplished at the system level where all available information about scheduled operations can be synthesized and optimized into a plan that incorporates all relevant resource and criticality information.

The definitions of the various information tiers are intentionally flexible, leaving room for interpretation on how a system is decomposed into its constituent subsystems, equipment, and component assets. Ultimately, the use and interpretation of these definitions can be application-specific to best suit user needs. For example, equipment may be composed into other equipment such as a machining center being decomposed into a constituent material handling robotic arm and a machine tool. Additionally, one could define digital equivalents of each of these levels that can be monitored and maintained in an analogous fashion to the physical assets presented in this paper. For simplicity, digital assets of this nature such as communications exchanges, cyber security protocols, operating systems, etc. are largely ignored with regards maintenance in this paper. The decisions and data requirements for each level are described in the sections below.

System level:

At the system level, we are primarily concerned with how maintenance decisions impact performance metrics such as throughput, cycle-time, and cost. From the maintenance perspective, important information is current and future equipment state and characterizing dynamic behaviors, such as

availability, capability, and capacity. Equipment standards, such as MTConnect, report whether a machine is available or unavailable (up/down) and possibly busy/idle. The second aspect related to planning and scheduling focuses on creating and incorporating predictions of future unavailability -- whether due to scheduled/planned maintenance tasks or estimated unplanned maintenance. There are several ways to incorporate information about expected availability into scheduling methods, dependent upon the specific available information [Vieira et al, 2003].

Accurate descriptions of equipment capability, capacity, and state are essential for operations management decision-making; primarily the scheduling of production and maintenance operations. This data collection enables identifying and characterizing bottleneck equipment (workstations and systems). Maintenance operations management decisions can then incorporate this information to determine maintenance priorities.

A brief summary of general system-level decision information and corresponding data requirements is presented below:

Classes of decisions:

Which production jobs can be assigned to a piece of equipment? Is it available and capable?
When should maintenance tasks be generated (planning) and performed (scheduling)?
How to prioritize maintenance jobs? When should jobs be performed and to what extent?
Which replacement components (or equipment) should be stocked in inventory?

Supporting Data:

- Equipment State (*Current and Future*)
  - Capability
  - Availability
  - Capacity
- Maintenance Queue
  - Priority/Criticality of tasks
  - Availability of resources (personnel, material, parts, etc.)
  - Impacts on throughput
  - Coordination of tasks (i.e. opportunity-driven priority)
- Production Queue
  - Priority/Criticality of tasks
  - Availability of resources (personnel, systems, material, parts, etc.)
  - Current buffer states

Equipment level:

At the equipment level, decisions are less about if the equipment *should* be run, and more about if it *could* be run. This centers around the determination of both the equipment health and availability, as well as capability and capacity; the determination if there is a configuration of the equipment that will allow for

safely fulfilling all requirements of the requested duty cycle. Synthesizing health information from the constituent components' health and making determinations of overall health as well as capacity will largely be diagnostic in nature. This information is also used to assess criticality of and assign priority to component maintenance requests.

Root cause investigations and determinations of potential solutions to the problem provide critical information for decision makers in the event of a failure. Additional information, such as the amount of required time/effort for a given solution, is also useful. Determinations of impacts on production and throughput would generally be made in context of the system level.

Another equipment level decision is the coordination between maintenance tasks. This is different than the assignment of criticality of the task, and focuses on the benefits from performing simultaneous maintenance activities --- effectively judging if the extra time spent at this piece of equipment will affect other maintenance jobs in the queue. These 'opportunity-driven' tasks can be aligned to minimize the total downtime of a given equipment.

Based on these decision requirements, much of the data and information collected around this level should support investigations relating to failure diagnostics, prediction, and prevention. The decision makers need to predict when the equipment will go down next, what will cause the failure, how long the corresponding maintenance will take, and how much additional time is required for the equipment and any associated process to resume normal operations. While a large portion of this information is fed from the component level, the prioritization and alignment of maintenance tasks can only be determined with the contextualization of the equipment level. Common examples of equipment level information come from raw and processed sensor signals from the component level (e.g., vibration data) as well as qualitative data in Maintenance Work Orders (MWOs) (e.g., description of the problem). The different decisions and supporting data is summarized below:

<u>Classes of decisions:</u>

- Should I operate this equipment?
    - Are all critical components in a state to allow safe completion of the planned duty cycle?
    - Can the equipment produce parts or perform services to the required minimum level of quality in its current state?
    - Can the equipment be in a different configuration to better meet requirements?
- What maintenance activities should be prioritized?
    - How critical are component level maintenance requests?
    - What, if any, is the equipment level relationship between diagnosed component degradation?
    - What is the criticality and time horizon of potential faults or failures?
    - Should any maintenance activities be grouped for simultaneous execution?
    - What are the maintenance solutions to prevent (or delay) failure and are these solutions cost effective to implement? If so, how long will it take to implement the solutions?
    - What is the root cause for an observed failure?
    - Has an observed failure happened before?

- What can be done to fix observed faults, failures, inefficiencies, or other problems and how long will it take to fix?
- How can I prevent similar problems in the future?

Supporting Data:

- Maintenance Work Order Data
  - Descriptions of previous faults/failures/etc. and corresponding solutions
  - Time spent on faults/failures and solutions
  - Technician(s) sent to solve fault/failure/etc.
  - Resources (e.g., parts, tools) used in addressing the fault/failure/etc.
- Equipment Data
  - Equipment manuals and schematics
  - Taxonomy of components in equipment
  - Population fault/failure rates
- Component States
  - Health information
  - Predicted faults/failures/etc.
  - Component level maintenance requests

Component level:

Component information monitoring focuses on two areas. The first area determines the current capabilities of the equipment: if and at what capacity or workload a component can be operated. The second relates to populating maintenance tasks via degradation monitoring, diagnostic fault cause analysis, and prediction of probable future states of the component. Some of this information needs to be contextualized at the equipment level, where understanding the relationships between components is essential, e.g., will this component hurt production efficiency, lower product quality, etc.? Even so, the bulk of the data/information regarding maintenance needs and even some decisions are collected/made at the component level.

The most simple information that can be captured is if the component currently able to operate, or if it is currently exhibiting a 'failure' condition. This can be a soft failure where the component is unable to perform at a minimal operating level or has deviated beyond allowable limits from expected behavior; or a hard failure, typically catastrophic, where the component is unable to function at any capacity. This class of information could loosely be categorized as a minimum state observation for the component. While this observation can provide decisions on 'go/no-go' scenarios, the amount of insight given is very minimal.

The next progression of component level decision information involves accessing the overall health and capacity of the component. This demand for extra information creates a corresponding amount of additional demand on the data required. There are an abundance of sensor types, algorithms, and models that can be used to access the current condition of a component.These assessments can be qualitative,

quantitative, definitive, or probabilistic, and are determined by either direct or indirect measurements about the system [Si 2011]. While different models and algorithms may impose different data needs for their construction, the general class of tools that measure or estimate the specific condition of a component all rely on some level of sensing capabilities. In some cases, certain condition monitoring tools can even benefit from higher level operational data, such as workload plans, maintenance activities, etc. A more in-depth breakdown of the data requirements for various modeling approaches (physics, rules-based, data driven, etc.) is beyond the scope of this brief paper, but will be addressed in future works. One metric useful at this information tier is the Current Life Consumed (CLC), which often corresponds to the current amount of degradation detected in the component, normally as a percentage of some failure threshold. This information tier encompases active monitoring of a component.

The component conditions assessment can also extend to predicting future states. This is accomplished by first knowing what are the current/future needs and expectations from the system, and second - given these expected stresses and demands - predicting the probable future condition and capacity of the component. These information types and analysis broadly fall under the category of prognostics. The 'prognosing' of future states can be longer term (e.g., for scheduling, maintenance, and planning), or shorter term that focuses exclusively on the current or immediately upcoming duty cycle. Some form of either definite or probabilistic operational plan must be defined or inferred for this analysis, which yields information beyond that available at the component level. This shows the relationships between multiple information levels and areas of a factory setting. Again, the specific tool or algorithm used for the prognostic assessment of the equipment will have specific needs of the component level information, either historical or current. A common metric for these types of analysis is a component's Remaining Useful Life (RUL). A brief summary of general component level decisions and corresponding data requirements is presented below:

<u>Classes of Decision:</u>

- Can a component, within a piece of equipment, be marked available for a requested operation?
    - Is the component currently occupied with some other task?
    - Is the component functioning or has it failed?
    - What are the state and time horizons of any currently identified incipient faults?
    - Can this component meet the current/future needs?
- What are the current or future maintenance actions required for this component?
    - Are any calendar-based or cycle-based preventive maintenance actions upcoming?
    - Are there any condition-based maintenance actions upcoming?
    - Are any corrective repairs needed?

<u>Supporting Data</u>

- Capacity Specifications
    - Possible configurations of component
    - Nominal work loads
- Condition assessment information
    - Human interrogators
    - Predictive models

    - Sensors
- Anticipated Future Performance
    - Planned duty cycles
    - Probabilistic modeling
- Maintenance planning
    - Needed resources
    - Maintenance work order data
    - OEM Maintenance recommendations

Summary and Conclusions

This paper discusses maintenance decision making and data needs at three levels: component level, equipment level, and system level. There is strong interplay between these levels with information flowing upwards from the component level LRUs, into a more contextualized equipment level of information and finally into system levels of information with pertinent decision support at every tier. The primary blocks of creating a maintenance action plan have different information needs at each of the three levels. A summary of example decisions and supporting questions for the various information levels is shown in Figure 2. These are not the only decisions or possible groupings of information that could be applied to a factory setting.

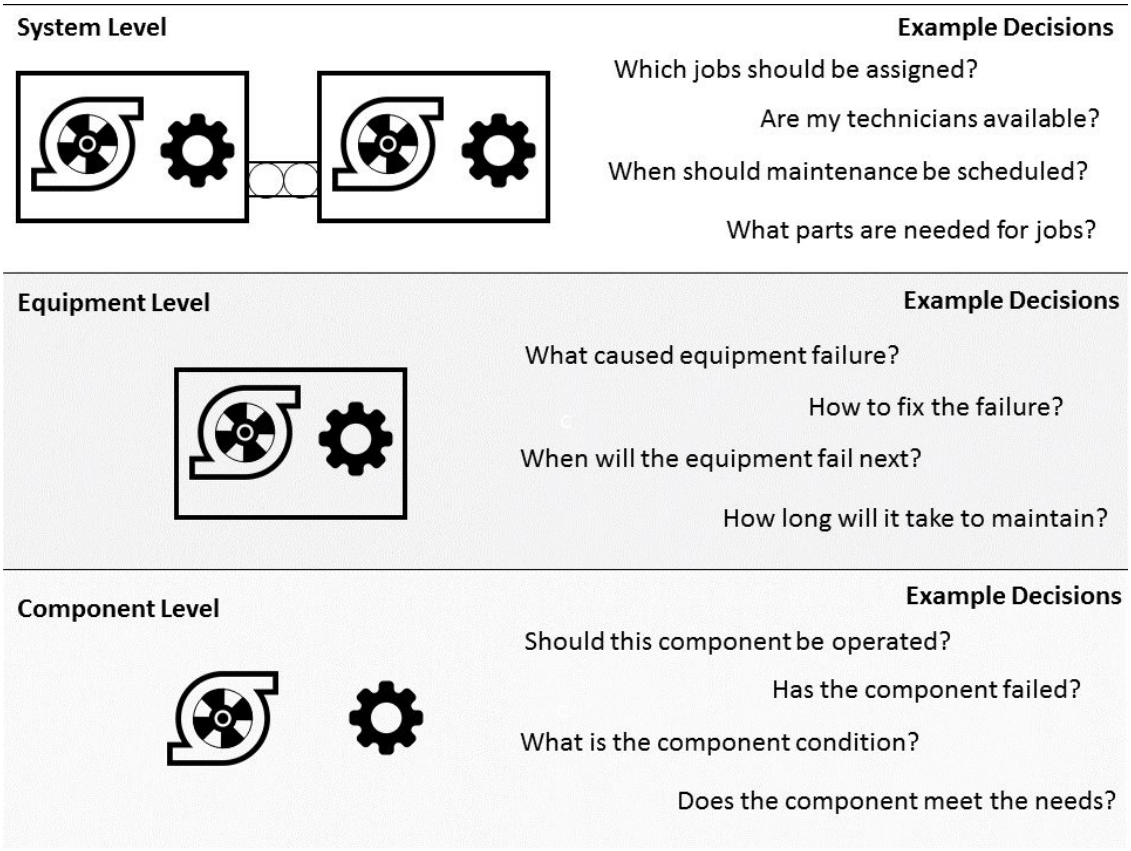


Figure 2: Maintenance Decisions and Corresponding Levels

The primary goal of this work is to present some common questions and decision support needs by way of information gathering. This is intended as the beginning steps to creating a working guide for industry practitioners to develop their own optimized information gathering and decision support network. Resources dedicated to gathering specific sources of information should be tailored to the explicit decision support needs of the managers, operators, and technicians. If an asset is not deemed mission critical, or is otherwise placed into a 'repair on failure' category, there is no need to invest in expensive data collection and storage tools. Conversely, if an asset is *highly* critical, there is a strong case for extensive monitoring and modeling tools to ensure that the asset rarely, if ever, experiences a failure. Other assets that could merit more monitoring tools are those that are creating large amounts of maintenance work orders, either calendar or condition-based. The extra monitoring and analysis could help to optimize the amount of maintenance and/or discover different operations or maintenance practices that could optimize the types of maintenance performed and thus reduce downtime.

The next steps in this work will look at a specific decision at the system level and create a framework for determining the correct data from both the equipment and component levels. This implementation will lead to a more appropriate roadmap of standards and research needed in this space.

NIST Disclaimer

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

References

ANSI QIF 3.0, October 5, 2018. http://qifstandards.org

Kong, Dongdong, Yongjie Chen, Ning Li. Gaussian Process Regression for Tool Wear Prediction. Mechanical Systems and Signal Processing. 2018. Vol, 104 (pp 556-574).

Kramer, Thomas R., Frederick M. Proctor, Elena R. Messina. The NIST RS274NGC Interpreter - Version 3. NIST Interagency/Internal Report (NISTIR) - 6556. August 01, 2000.

Lee, Gil-Yong & Kim, Mincheol & Quan, Yingjun & Kim, Min-Sik & Joon Young Kim, Thomas & Yoon, Hae-Sung & Min, Sangkee & Kim, Dong-Hyeon & Mun, Jeong-Wook & Woo Oh, Jin & Gyu Choi, In & Kim, Chung-Soo & Chu, Won-Shik & Yang, Jinkyu & Bhandari, Binayak & Lee, Choon-Man & Ihn, Jeong-Beom & Ahn, Sung-Hoon. (2018). Machine health management in smart factory: A review. Journal of Mechanical Science and Technology. 32. 987-1009. 10.1007/s12206-018-0201-1.

Li, Rui, Wim J.C. Verhagen, and Richard Curran. A Functional Architecture of Prognostics and Health Management Using a Systems Engineering Approach. European Conference of the Prognostics and Health Management Society, 2018.

MTConnect Standard. 2018 MTConnect Institute. https://www.mtconnect.org/

Sackett, J., P & F. Al-Gaylani, M & Tiwari, Ashutosh & Williams, D. (2006). A review of data visualization: Opportunities in manufacturing sequence management. Int. J. Computer Integrated Manufacturing. 19. 689-704. 10.1080/09511920500504578.

Si, Xiao-Sheng, Wenbin Wang, Chang-Hua Hu, Dong-Hua Zhou, Remaining useful life estimation – A review on the statistical data driven approaches, European Journal of Operational Research, Volume 213, Issue 1, 2011, Pages 1-14, ISSN 0377-2217, https://doi.org/10.1016/j.ejor.2010.11.018.

Sexton, Thurston, Michael P. Brundage, Melinda Hodkiewicz, Thomas Smoker. Benchmarking for Keyword Extraction Methodologies in Maintenance Work Orders. 2018 Annual Conference of the Prognostics and Health Management Society. September 24-27, 2018. Philadelphia, PA.

Sharp, Michael, Brian Weiss. Hierarchical modeling of a manufacturing work cell to promote contextualized PHM information across multiple levels Manufacturing Letters, Volume 15, Part A, January 2018, Pages 46-49.

United States, Congress, “ISO 10303-11 Industrial Automation Systems and Integration: Product Data Representation and Exchange.” ISO 10303-11 Industrial Automation Systems and Integration: Product Data Representation and Exchange, ISO, 1994

Vieira, G. E., Herrmann, J. W., & Lin, E. (2003). Rescheduling manufacturing systems: a framework of strategies, policies, and methods. Journal of scheduling, 6(1), 39-62.

Sub Topic: Laminar Flames

11th U. S. National Combustion Meeting
Organized by the Western States Section of the Combustion Institute
March 24–27, 2019
Pasadena, California

R-152a/air and R-134a/oxygen constant volume spherical flame burning velocity measurements

Robert R. Burrell, Michael J. Hegetschweiler, Donald R. Burgess Jr., Jeffrey A. Manion, Valeri I. Babushok, Gregory T. Linteris*

Energy and Environment Division, National Institute of Standards and Technology, Gaithersburg, Maryland, 20899, USA

**Corresponding Author Email: Linteris@nist.gov*

Abstract: Many presently used refrigerants are non-flammable but are being phased out due to concerns about their high global warming potential (GWP). Replacements with low GWP exist but tend to be flammable with a maximum burning velocity in air between 1 cm/s and 40 cm/s, depending on molecular structure. Flammable refrigerants are a rising challenge for industry, which can benefit from predictive tools for ranking refrigerant flammability based on fundamentals. This work reports experimental burning velocities via pressure rise measurements in a constant volume spherical chamber interpreted with the aid of a thermodynamic spherical flame model. Flames of R-152a/air and R-134a/oxygen mixtures over a range of equivalence ratios provide experimental burning velocities for unburned gas conditions at 298 K and 0.101 MPa, and at 375 K and 0.253 MPa. This work supports the development of validated and optimized kinetic models for the combustion of refrigerants at conditions relevant to fire safety.

Keywords: Refrigerant Flammability; Spherical Flame; 1,1 Difluoroethane; 1,1,1,2-Tetrafluoroethane

1. Introduction

The world is phasing down use of refrigerants with high global warming potential (GWP). One way to limit GWP is to use refrigerants with higher reactivity and shorter atmospheric half-life. Higher reactivity also often implies greater flammability in air. Flammable refrigerants are a rising challenge for the heating, ventilation, air-conditioning, and refrigeration industry, which can benefit from tools for predicting flammability characteristic based on fundamentals. Efforts are underway at the National Institute of Standards and Technology (NIST) to develop a comprehensive chemical kinetic model for refrigerant combustion validated against fundamental burning velocity (S_u^0) data. The definition for S_u^0 applies to flames that are freely propagating, laminar, one-dimensional, planar, adiabatic, and stretch free. In laboratory flames subject to flame stretch and/or heat loss, the burning velocity (S_u) can deviate from the ideal S_u^0.

Combustion kinetic models are hierarchical in nature; reactions for small chemical species form a subset of those for larger species. For example, methane combustion reactions are a subset of

Burrell, Robert; Linteris, Gregory; Burgess Jr., Donald; Hegetschweiler, Michael; Manion, Jeffrey; Babushok, Valeri.
"R-152a/air and R-134a/oxygen constant volume spherical flame burning velocity measurements."
Paper presented at 11th U.S. National Meeting of the Combustion Institute, Pasadena, CA, United States. March 24, 2019 - March 27, 2019.

those for ethane and higher alkanes. The authors previously developed a kinetic model for R-32 (difluoromethane) combustion from a critical evaluation of reaction rates and validated it against S_u^0 measurements [1-3]. The authors now turn attention to the two-carbon refrigerants R-134a (1,1,1,2-tetrafluoroethane) and R-152a (1,1-difluoroethane). R-134a is a widely used working fluid for stationary refrigeration and automobile air-conditioning systems. It has a low H/F atom ratio of ½, resulting in low overall reactivity and non-flammability at typical atmospheric conditions. There are no known S_u^0 measurements for flames of a single-component R-134a fuel. It has a high GWP (1300 times that of CO_2) and its use is being phased down. R-152a is not widely used as a refrigerant, but it offers similar thermodynamic cycle performance as R-134a. It has an H/F ratio of 2 which results in moderate flame reactivity and a much lower GWP (140 times that of CO_2). Measurements by Takizawa et al. [4] and Moghaddas et al. [5] using constant volume method spherical flames indicate a maximum R-152a/air S_u^0 at standard conditions near 23 cm/s.

The goal of this work is to further support development and validation of the NIST refrigerant combustion kinetic model with S_u^0 data for R-134a/O_2 and R-152a/air mixtures. This is achieved by taking pressure rise measurements of outwardly propagating spherical flames in a constant volume spherical chamber. S_u is inferred from the measured pressure history using a thermodynamic model to determine the relationship between chamber pressure and flame radius. The influence of initial transients, flame instabilities, and flame-wall interactions on data reduction are investigated. Data not affected by these influences are used to produce experimental S_u^0 values for R-134a/O_2 mixtures over a range of fuel-oxidizer equivalence ratios (ϕ) from $0.6 \leq \phi \leq 1.2$ at (298 K, 0.101 MPa) and for R-152a/air mixtures over $0.8 \leq \phi \leq 1.3$ at (298 K, 0.101 MPa) and (375 K, 0.253 MPa).

2. Experimental Methods

Measurements were performed using the 15.24 cm inner diameter spherical chamber described previously [3, 6, 7]. Reactants were R-134a, R-152a (99.5 %), O_2 (99.5 %), and house filtered/dried air. The chamber was evacuated to below 67 Pa for 5 minutes, then reactants were added by partial pressure as measured with an absolute pressure transducer (Omega PX811) calibrated to an in-house reference (Baratron 627D) until the desired ϕ was achieved. R-134a/O_2 mixtures were prepared at an initial state T_i = 298 K and P_i = 0.1013 MPa for $0.6 \leq \phi \leq 1.2$. R-152a/air mixtures were prepared for multiple initial states with T_i = 298 K and P_i = 0.0880 MPa, 0.1013 MPa, and 0.1147 MPa. Variability in T_i was $\pm$ 2 K and in P_i was $\pm$ 0.0001 MPa. Mixtures were given 5 minutes to settle then centrally ignited by a spark powered by either a 5 nF or 10 nF capacitor bank charged to between 6 kV to 14 kV. Estimated ignition energies range between 0.31 mJ to 5.6 mJ [8]. Pressure rise in the chamber caused by the outward expansion of a spherical flame was recorded using a piezoelectric sensor at 10 kHz in R-134a/O_2 mixtures and 5 kHz in R-152a/air mixtures. During post-processing, the effective data rate was set to 1.667 kHz to reduce the relative effect of measurement noise. A previous uncertainty analysis of the present system [7] indicates the maximum S_u uncertainty for a single point measurement is $\pm$12 % (2σ) and occurs in off-stoichiometric mixtures, which is about $\pm$1 cm/s in R-134a/O_2 flames and $\pm$2 cm/s in R-152a/air flames.

Burrell, Robert; Linteris, Gregory; Burgess Jr., Donald; Hegetschweiler, Michael; Manion, Jeffrey; Babushok, Valeri.
"R-152a/air and R-134a/oxygen constant volume spherical flame burning velocity measurements."
Paper presented at 11th U.S. National Meeting of the Combustion Institute, Pasadena, CA, United States. March 24, 2019 - March 27, 2019.

3. Data Reduction

Given that only the pressure (P) vs. time (t) history is measured, determining the corresponding S_u requires modeling to relate the flame radius (R_f) to P. This was achieved using a two-zone, thermodynamic, spherical flame model [6, 9, 10], with thermodynamic data for 13 species (O_2, N_2, CO_2, HF, CO, H_2O, H_2, CF_4, CF_2O, OH, O, F, and H) taken from the NASA CEA2 database [11, 12], and heats of formation for R-134a and R-152a of -910 kJ/mol and -512 kJ/mol, respectively, taken from the NIST refrigerant combustion kinetic model. The two-zone model divides the chamber into adiabatic burned and unburned zones, each with uniform temperature and composition separated by a smooth, spherical, and infinitesimally thin flame front. No reactions occur in the unburned gas and the burned gas is at chemical equilibrium. The pressure is spatially uniform but rises in time due to isentropic compression from the expanding flame.

The model divides the spherical volume of reactants at initial state (T_i, P_i) into a series of shells spaced to correspond with the distance the flame propagates over a segment of the measured pressure. A shell is burned by computing the equilibrium state of the unburned reactants for constant P and enthalpy. This increases the burned shell volume and temperature (T_b), but the unburned shell volume and temperature (T_u) are unchanged. Both volumes are simultaneously isentropically compressed to match the chamber volume, which increases T_u and P. In practice, this is achieved by fixing P to the measured values and iteratively solving for T_b and the burned gas mass fraction (χ_b). The flame radius (R_f) is calculated as $R_f = R_w[1-(1-\chi_b)(P_i/P)^{\frac{1}{\gamma_u}}]^{\frac{1}{3}}$, where R_w is the chamber wall radius and γ_u is the unburned gas specific heat ratio. Finally, the experimental S_u values were calculated:

$$S_u = \frac{R_w}{3}\left(\frac{R_w}{R_f}\right)^2\left(\frac{P_i}{P}\right)^{\frac{1}{\gamma_u}}\left(\frac{\mathrm{d}\chi_b}{\mathrm{d}t}\right) \qquad \text{Eqn. 1}$$

which depends on measured $P(t)$ and modeled components $\gamma_u(P)$ and $\chi_b(P)$. The resulting S_u evolves in time along an isentrope in (T_u, P) space unique to the initial state of the reactants.

The data for a given ϕ were fit to a power law surface, $\widehat{S_u} = S_{u,ref}(T_u/T_{u,ref})^a(P/P_{ref})^b$, where $S_{u,ref}$ is the S_u value of the fit surface at reference unburned gas conditions ($T_{u,ref}$, P_{ref}) and the exponents a and b are the temperature and pressure dependence of S_u as determined by the fitting process. An example $\widehat{S_u}$ surface is shown for R-152a/air at ϕ = 1.1 in Fig. 1 with the corresponding S_u data (orange circles) also shown projected onto each coordinate plane (black dots). Two (T_u P) states are marked (blue squares) in Fig. 1 at which experimental S_u^0 values are provided for R-152a/air mixtures, (298 K, 0.101 MPa), which results from extrapolation, and (375 K, 0.253 MPa) obtained through interpolation.

Choosing appropriate (T_u, P) conditions for interpolation or extrapolation is dictated by the experimental range. Consider the (T_u,P) plane in Fig. 1 which shows the unburned gas histories

Burrell, Robert; Linteris, Gregory; Burgess Jr., Donald; Hegetschweiler, Michael; Manion, Jeffrey; Babushok, Valeri.
"R-152a/air and R-134a/oxygen constant volume spherical flame burning velocity measurements."
Paper presented at 11th U.S. National Meeting of the Combustion Institute, Pasadena, CA, United States. March 24, 2019 - March 27, 2019.

of R-152a/air flames at three initial pressures. A wide range of conditions can be selected that are bounded by the experimental isentropes, e.g., via interpolation to (375 K, 0.253 MPa). On the other hand, R-134a/air S_u curves (not shown) were prepared at only one initial pressure. This situation is analogous to only using one of the three R-152a/air (T_u, P) histories to fit $\widehat{S_u}$. A single S_u curve is two-dimensional and does not constrain the surface fit in the direction orthogonal to the curve. However, the fit remains well-constrained in the direction tangential to the S_u curve, thus interpolation or extrapolation can still be performed along the isentropes, e.g., via extrapolation to the initial condition (298 K, 0.101 MPa).

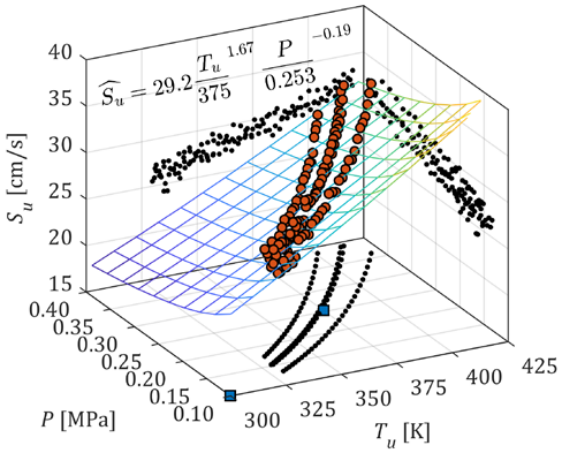


Fig. 1: Comparison of fitted $\widehat{S_u}$ surface (mesh) and experimental S_u values (orange circles) for R-152a/air at ϕ = 1.1 with the locations of conditions used to report S_u^0 (blue squares) and projections of S_u data onto coordinate planes (black dots).

4. Results and Discussion

Determining the proper range of flame data for fitting surfaces is a matter of interpretation. Fig. 2 shows examples of experimental S_u data for R-134a/O_2 mixtures projected onto the (T_u, S_u) plane for clarity. Starting near the initial condition of $T_u = T_i = 298$ K, the flame radius is small and the S_u noise is initially large and then decays. As the flame sphere grows, T_u increases, noise reduces, and an increasing S_u trend forms for $T_u > 310$ K, which become quasi-linear for $T_u >$ 320 K. For $T_u > 375$ K, there is apparent flame acceleration initiating at lower T_u for richer mixtures. The reason for this is believed to be the diffusive-thermal instability [13] known (in rich hydrocarbon/air flames) to cause a non-smooth, cellular flame that leads to enhanced overall mass burning rates. In all cases, S_u decreases as the flame becomes large enough to quench near the chamber walls.

The characteristics of experimental S_u values in R-152a/air flames, shown in Fig. 3, are broadly similar to those identified for R-134a/O_2 flames. There is an initial period of high measurement

Burrell, Robert; Linteris, Gregory; Burgess Jr., Donald; Hegetschweiler, Michael; Manion, Jeffrey; Babushok, Valeri.
"R-152a/air and R-134a/oxygen constant volume spherical flame burning velocity measurements."
Paper presented at 11th U.S. National Meeting of the Combustion Institute, Pasadena, CA, United States. March 24, 2019 - March 27, 2019.

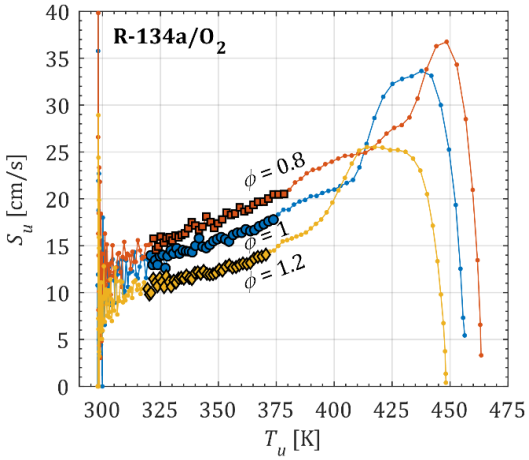


Fig. 2: Experimental S_u values for initial conditions (298 K, 0.101 MPa).

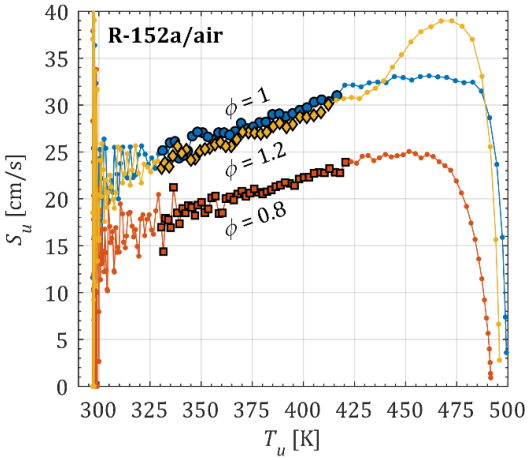


Fig. 3: Experimental S_u values for initial conditions (298 K, 0.101 MPa).

noise followed by quasi-linear $S_u(T_u)$ response culminating with S_u going to zero. Evidence for cellular instability was observed at ϕ = 1.2 but not in leaner mixtures. The reasons for the reduced occurrence of instabilities in R-152a/air mixtures are not presently clear. One possibility is lower flame temperatures for R-152a/air mixtures than R-134a/O_2. For example, at $P = 4P_i$ for the ϕ = 1 curves shown in Figs. 2 and 3, the R-152a/air unburned mixture is near 420 K with a calculated adiabatic flame temperature of 2310 K while the R-134a/O2 reactants are only near 370 K with a higher adiabatic flame temperature above 2400 K. Higher flame temperatures drive greater unburned-to-burned mass density ratios across the flame front, promoting cell formation via hydrodynamic instability [13].

Taking the raw S_u data characteristics into consideration, some guidelines can be established for acceptable S_u data. Initial transients, cellular instabilities, and flame-wall interactions should be avoided by looking for a quasi-linearly increasing $S_u(T_u)$ trend at intermediate T_u. It is convenient to formalize the conditions in terms of P, which is the only measured variable. The initial transients were avoided by eliminating data with $P < 1.5P_i$. Cellular instabilities and flame-wall interactions were avoided by eliminating data with $P > 4P_i$. Data conforming to these pressure conditions are denoted by symbols with black borders in Fig. 2 and Fig. *3*.

Experimental S_u^0 values derived from $\widehat{S_u}$ fits are shown in Fig. 4 for R-134a/O_2 flames at (298 K, 0.101 MPa). A maximum S_u^0 of 10.8 cm/s was observed near $\phi = 0.7$ (uncertainties, 1σ, are due to the surface fit and extrapolation). To the authors' knowledge, these are the first S_u^0 measurements reported in flames using a single-component R-134a fuel. Due to the relatively low burning velocities, the present S_u^0 values are likely to be influenced by burned gas thermal radiation, which has not presently been addressed. The absence of N_2 in the oxidizer increases the concentration of the emitting/absorbing species. Radiation corrections in R-32/air flames [3]

Burrell, Robert; Linteris, Gregory; Burgess Jr., Donald; Hegetschweiler, Michael; Manion, Jeffrey; Babushok, Valeri.
"R-152a/air and R-134a/oxygen constant volume spherical flame burning velocity measurements."
Paper presented at 11th U.S. National Meeting of the Combustion Institute, Pasadena, CA, United States. March 24, 2019 - March 27, 2019.

increased S_u^0 by about 15 %. Future work will quantify the impact of thermal radiation in these flames.

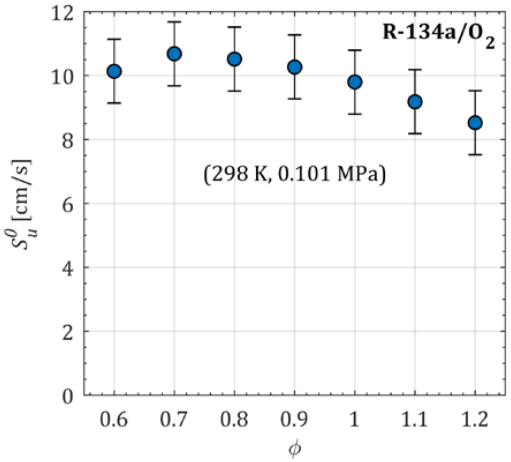


Fig. 4: Experimental S_u^0 values for R-134a/O_2 mixtures.

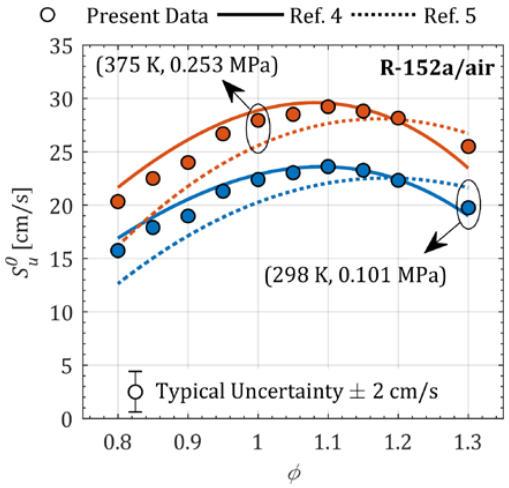


Fig. 5: Experimental S_u^0 values for R-152a/air mixtures with literature values.

Experimental S_u^0 values derived from $\widehat{S_u}$ fits are shown in Fig. 5 for R-152a/air flames at and (375 K, 0.253 MPa) along with correlations from measurements by Takizawa et al. [4] and Moghaddas et al. [5]. A maximum experimental S_u^0 of 24 cm/s was observed near $\phi = 1.1$ at (298 K, 0.101 MPa), increasing to 29 cm/s at (375 K, 0.253 MPa). The magnitude of S_u^0 is less than seen in hydrocarbon/air flames and there is less variation of S_u^0 with ϕ, although the shape is similar. The similarities to hydrocarbon/air flames may be attributed to the refrigerant H/F ratio significantly above unity. Although burned gas thermal radiation corrections have not been made, the relatively high S_u^0 values for R-152a/air mixtures suggest that the correction will be small. Neither Refs. 4 nor 5 applied radiation corrections, so a straightforward comparison to the present data is possible. Generally, the present burning velocities compare well with the experimental correlations from Ref. 4, but are slightly slower on the lean side and faster on the rich side. The experimental correlation values from Ref. 5 are lower in maximum value, and the S_u^0 vs ϕ shape is shifted toward richer mixtures compared to both Ref. 4 and the present S_u^0 values.

5. Conclusions

Experimental burning velocities of R-152/air and R-134a/O_2 mixtures were obtained by measuring the pressure rise caused by a spherically expanding flame in a constant volume spherical chamber. Burning velocities were extracted from the measured pressure traces using a two-zone thermodynamic model to obtain the relationship between the chamber pressure and the flame radius. Data taken at early times were deemed unreliable due to high measurement noise and data at late times were affected by flame-wall interactions. For richer flames, a rapid burning velocity increase at higher temperatures was believed to be due to the onset of flame instabilities. These data were eliminated when deriving experimental burning velocities. Stretch-free fundamental

Burrell, Robert; Linteris, Gregory; Burgess Jr., Donald; Hegetschweiler, Michael; Manion, Jeffrey; Babushok, Valeri.
"R-152a/air and R-134a/oxygen constant volume spherical flame burning velocity measurements."
Paper presented at 11th U.S. National Meeting of the Combustion Institute, Pasadena, CA, United States. March 24, 2019 - March 27, 2019.

burning velocities of R-134a/O_2 flames were reported for equivalence ratios from 0.6 to 1.2 at 298 K and 0.101 MPa. The data indicate a peak fundamental burning velocity of about 10 cm/s for lean equivalence ratios near 0.7. Burning velocities of R-152a/air flames were reported for equivalence ratios of 0.8 to 1.3 at 298 K and 0.101 MPa and 375 K and 0.253 MPa. The data indicate a peak fundamental burning velocity at 298 K and 0.101 MPa of about 24 cm/s near an equivalence ratio of 1.1. Future work will quantify the influence of flame stretch and burned gas thermal radiation on measurements.

6. Acknowledgements

This work was supported by the Buildings Technologies Office of the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy under contract no. DO-EE0007615 with Antonio Bouza serving as Project Manager. Dr. Burrell was supported by a NIST National Research Council postdoctoral research associateship.

7. References

[1] D. R. Burgess Jr, J. A. Manion, R. R. Burrell, V. I. Babushok, M. J. Hegetschweiler and G. T. Linteris, Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures, Spering Meetings of the Eastern States Section of the Combustion Institute (2018).

[2] D. R. Burgess Jr, J. A. Manion, R. R. Burrell, V. I. Babushok, M. J. Hegetschweiler and G. T. Linteris, Validated model for burning velocities of R-32/O2/N2 Mixtures over a wide range of conditions, 11th US National Combustion Meeting (2019).

[3] R. R. Burrell, J. L. Pagliaro and G. T. Linteris, Effects of stretch and thermal radiation on difluoromethane/air burning velocity measurements in constant volume spherically expanding flames, Proceed. Combust. Inst. (2019) 4231-4238.

[4] K. Takizawa, A. Takahashi, K. Tokuhashi, S. Kondo and A. Sekiya, Burning velocity measurement of fluorinated compounds by the spherical-vessel method, Combust. Flame (2005) 298-307.

[5] A. Moghaddas, C. Bennett, E. Rokni and H. Metghalchi, Laminar burning speeds and flame structures of mixtures of difluoromethane (HFC-32) and 1,1-difluoroethane (HFC-152a) with air at elevated temperatures and pressures, HVAC&R Res. (2014) 42-50.

[6] J. L. Pagliaro, G. T. Linteris, P. B. Sunderland and P. T. Baker, Combustion inhibition and enhancement of premixed methane-air flames by halon replacements, Combust. Flame (2015) 41-49.

[7] J. L. Pagliaro, G. T. Linteris and V. I. Babushok, Premixed flame inhibition by C2HF3Cl2 and C2HF5, Combust. Flame (2016) 54-65.

Burrell, Robert; Linteris, Gregory; Burgess Jr., Donald; Hegetschweiler, Michael; Manion, Jeffrey; Babushok, Valeri.
"R-152a/air and R-134a/oxygen constant volume spherical flame burning velocity measurements."
Paper presented at 11th U.S. National Meeting of the Combustion Institute, Pasadena, CA, United States. March 24, 2019 - March 27, 2019.

[8] J. L. Pagliaro, Inhibition of laminar premixed flames by Halon 1301 Alternatives, Thesis, Department of Fire Protection Engineering, University of Maryland, College Park, 2015.

[9] M. Metghalchi and J. C. Keck, Laminar burning velocity of propane-air mixtures at high temperature and pressure, Combust. Flame (1980) 143-154.

[10] P. G. Hill and J. Hung, Laminar Burning Velocities of Stoichiometric Mixtures of Methane with Propane and Ethane Additives, Combust. Sci. Technol. (1988) 7-30.

[11] S. Gordon and B. J. McBride, Computer program for calculation of complex chemical equilibrium compositions and applications, Report No. NASA Reference Publication 1311, NASA Glenn Research Center, Cleveland, OH, USA, 1996.

[12] NASA Chemical Equilibrium Solver with Applications (CEA), NASA Glenn Research Center, Cleveland, OH, USA, https://www.grc.nasa.gov/www/CEAWeb/, updated: Feb. 4, 2016.

[13] M. Matalon, Flame dynamics, Proceed. Combust. Inst. (2009) 57-82.

Principles for Designed-In Security and Privacy for Smart Cities

Corey Dickens †
Dakota Consulting, Inc.
Silver Spring, MD, USA
corey.dickens@nist.gov

Paul Boynton †
National Institute of Standards and Technology
Gaithersburg, MD, USA
paul.boynton@nist.gov

Sokwoo Rhee †
National Institute of Standards and Technology
Gaithersburg, MD, USA
sokwoo.rhee@nist.gov

ABSTRACT

This paper presents the design and implementation of a process for an exploratory study that identifies a set of principles for designed-in security and privacy for smart city projects from among Global City Teams Challenge (GCTC) – Smart and Secure Cities and Communities Challenge (SC3) participants. The study was conducted based on information from the National Institute of Standards and Technology (NIST) GCTC Action Clusters database and interactions with the project teams. A research process was developed and implemented, comprising the following three steps:

(1) Investigate project descriptions created by the project leads on the NIST GCTC database and other public sources;

(2) Gather additional input from volunteer GCTC collaborators; and

(3) Identify a set of governing principles commonly shared by examples of GCTC projects.

Based on the outcomes of this process, a set of common principles has been identified that enable designed-in security and privacy considerations among the projects: specific technology usage, implementation of a cybersecurity management process and framework, and cybersecurity expertise and public-private partnerships. Characteristics of planning and implementation of security and privacy considerations from four example GCTC projects are described and analyzed in detail to illustrate the process.

†Smart Grid and Cyber-Physical Systems Program Office, Engineering Laboratory, National Institute of Standards and Technology, U.S. Department of Commerce

CCS CONCEPTS

• Smart City • Cybersecurity Framework • Cyber-Physical Systems • Internet of Things • Network Infrastructure

KEYWORDS

Smart city, Cybersecurity, Cyber-Physical Systems, IoT, Infrastructure, Privacy, Resiliency

ACM Reference format:

Corey Dickens, Paul Boynton, and Sokwoo Rhee. 2019. Principles for Designed-In Security and Privacy for Smart Cities. In *Proceedings of ACM CPS-Week. ACM, New York, NY, USA, 5 pages.*
https://doi.org/10.1145/3313237.3313300

1 Introduction

Cities and communities across the US and internationally are attempting to implement community-scale smart systems to improve city management efficiency, increase infrastructure resilience, and provide more convenient and secure services for their constituents [1][2]. These smart city and community initiatives can be complicated projects composed of many different technical domains, including information technology, communication networks, computer analytics, sensors, power electronics, and control systems. These systems and technologies need to interact and communicate in a secure and privacy-protecting manner to effectively operate and meet the needs of city and community stakeholders in terms of privacy and information integrity [3].

Due to numerous interoperability complexities and the lack of in-house cybersecurity expertise, many city and communities are initiating smart city projects in which security and privacy are addressed after initial project development and implementation (rather than considered as an integral component designed-in at the beginning of the project). Although this approach may seem simpler at the beginning, it does not guarantee systematic and holistic implementation of security and privacy that works best for the given smart city project. Moreover, additional patchworks for

Dickens, Corey; Boynton, Paul; Rhee, Sokwoo.
"Principles for Designed-In Security and Privacy for Smart Cities."
Paper presented at Cyber-Physical Systems and Internet-of-Things Week (CPS-IoT Week 2019), Montreal, Canada. April 15, 2019 - April 18, 2019.

security and privacy as an "afterthought" may lead to unintended errors and suboptimal performance of the implemented system. Therefore, it is important to identify a set of governing principles for designed-in security and privacy for smart city projects and take them into account at the beginning of project planning. This paper describes our work in identifying a common set of principles for security and privacy consideration that offer best practices and guidelines that could be used by communities as examples and a template for cybersecurity planning and integration for smart city projects. To illustrate the process, the paper describes several examples of existing smart city projects that are designed and planned from a perspective of trustworthiness, privacy, and security and provide examples that other cities and communities can consider in developing their own plans using available tools such as National Institute of Standards and Technology's (NIST's) Cyber-Physical Systems (CPS) Framework [4].

2 Methodology

The methodology to identify a common set of designed-in security and privacy principles followed an investigation of Global City Teams Challenge (GCTC) project descriptions in the NIST GCTC database [5] and discussions with GCTC collaborators. The process used to identify the principles from the technical use cases can be summarized as follows:

(1) Investigate project descriptions created by the project leads on the NIST GCTC database and other public sources;

(2) Gather additional input from volunteer GCTC collaborators; and

(3) Identify a set of governing principles commonly shared by examples of GCTC project.

2.1 Investigating GCTC Database Project Descriptions and other Public Sources

The GCTC is a program initiated and led by NIST, and implemented in partnership with other U.S. federal agencies including the Department of Homeland Security Science and Technology Directorate (DHS S&T). The program was established in 2014 as an outgrowth of the successful SmartAmerica Challenge launched in 2013 [6] to encourage development of replicable and scalable models for incubation and deployment of interoperable, secure, standard-based solutions using advanced technologies, such as Internet of Things (IoT) and CPS, and to demonstrate their measurable benefits in cities and communities. A cornerstone of the program is to incentivize public–private collaborations to develop sustainable integration models that can be used globally by interested community stakeholders [7]. Cities and communities collaborate with technology innovators to develop smart city projects, called Action Clusters, in relation to a chosen sector, with an aim to manage and control city resources in a secure and efficient manner (Figure 1). GCTC Action Clusters cover a wide range of application sectors that cover many infrastructures. Table

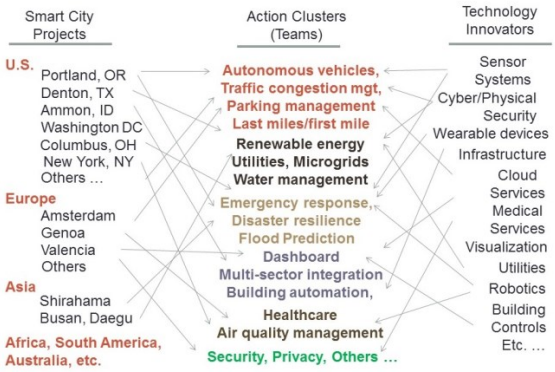


Figure 1. GCTC Action Clusters

1 lists several application sectors covered by GCTC Action Clusters.

GCTC participants are ideal for this type of exploratory case study because they are already engaged in smart city projects and detailed in the GCTC database. Therefore, examples of smart city projects were identified from among the past participants of GCTC and additional information was obtained from other public sources such as the GCTC Wiki [8].

Table 1. GCTC Action Cluster Sectors

Application Areas
Transportation
Public Safety
Utility (Energy, Water, and Environment)
Data Governance and Exchange
Wireless
Education
Agriculture and Rural

2.2 Gathering Additional Input from GCTC Collaborators

The GCTC project descriptions created by the project leads in the database provided an overview of participants' projects, but in most cases, they offered limited information about cybersecurity integration. Therefore, conversations were needed with participants to fill in the gaps for cybersecurity details that were not discussed in descriptions posted in the NIST database.

To fill in the information gaps about cybersecurity integration, remote meetings were set up with several GCTC teams to explain

Dickens, Corey; Boynton, Paul; Rhee, Sokwoo.
"Principles for Designed-In Security and Privacy for Smart Cities."
Paper presented at Cyber-Physical Systems and Internet-of-Things Week (CPS-IoT Week 2019), Montreal, Canada. April 15, 2019 - April 18, 2019.

the goal of the analysis and gather additional information on their projects.

2.3 Analysis of the Projects and Identifying a Set of Principles for Designed-in Security and Privacy Implementation

The GCTC collaborative partners described their projects in terms of cybersecurity integration as an initial design priority. Summaries of the discussions and project reviews were developed detailing specific information pertinent to relevant concepts and topics for designed-in security and privacy for smart city projects. From the detailed descriptions collected through the investigation of project descriptions on the NIST GCTC database, and one-on-one conversations with the GCTC collaborators, a set of common considerations for designed-in security and privacy could be identified.

3 Smart City Project Use Cases

Four projects are described in this paper as examples of uses cases. The four projects consist of:

(1) Secure Smart Lights / Sensors project, San Leandro, California, USA;

(2) Smart Trip Las Vegas: Safer and Connected Transportation, Las Vegas, Nevada, USA;

(3) Building Portfolio Cyber-Secure, Real-Time Utility Data Integration, Pittsburgh, Pennsylvania, USA; and

(4) Underground Infrastructure Sensing and Mapping for Smart Maintenance, Sustainability, Usage and Resilience, Burlington Vermont, USA.

3.1 Secure Smart Lights/Sensors, San Leandro, California

The San Leandro project is a utility/energy sector project that is comprised of a secure controllable network of efficient street lights providing reduced energy consumption and safety [9]. The lighting solution offers autonomous or scheduled illumination dimming that can be controlled from the district level down to a specific light node.

The core team, a public-private partnership, consists of a government partner, the City of San Leandro (CSL), and two technology private companies. CSL decided to outsource the project for a developed turn-key solution. The companies collaborated to provide the cybersecurity expertise and the LED based smart lighting solution. The cybersecurity solution adopted by the city is based on the concept of “moving target data protection” designed to render common attacks intractable as it removes contextual cues while increasing obfuscation via various distributed computation techniques [9].

The cybersecurity concerns are handled using the Information Security Maturity Model (ISM3) framework with the Open Source Security Information Management (OSSIM) tool [10][11]. The OSSIM platform implemented provides intrusion detection and prevention.

3.2 Smart Trip Las Vegas: Safer Connected Transportation, Las Vegas, Nevada

The Las Vegas project is a transportation sector project that uses a technology that provides fleet vehicles the capability to gather intelligence about traffic and pedestrian issues to increase pedestrian safety. The project aims to use connected car technologies to provide actionable messaging to drivers on road conditions while adopting open standard vehicle communication protocols based on a consortia-developed technology enabling faster deployment among vehicle manufacturers. The system establishes secure, reliable, two-way wireless connectivity between the in-vehicle unit and city-hosted servers so that vehicle data flows and actionable alerts return to inform the driver. The project also develops analytics and trend reports that support city/region/state transportation planning decisions for the welfare of citizens and visitors [12].

The team, a public-private partnership, consists of two government entities—the City of Las Vegas (CLV) and the Regional Transportation Commission (RTC) of Southern Nevada—and a technology private company.

The two municipal entities provided management and the fleet vehicles. The technology company developed the turnkey solution in terms of software, hardware, integration, and security. The technology was integrated without a cyber-based management plan, but with a security framework leveraged by the private partner and security protocols that were used in the technology.

3.3 Building Portfolio Cyber-Secure, Real Time Utility Data Integration, Pittsburgh, Pennsylvania

The Pittsburgh project is leveraged from an energy efficiency project in the U.S. Department of Energy (DOE) sponsored energy efficiency hub. The foundation of the project focused on plug load technology coupled with utility portfolio data management for energy management and reduction for commercial buildings. Based on visualizations and artificial intelligence algorithms developed at the Carnegie Mellon University (CMU) School of Architecture (SOA), the solution is a cloud-based building information system that can integrate with utility and building data to provide visualizations, benchmarking and recommendations for building portfolio energy efficiencies and indoor environmental quality. This project also plans to include the first independent Smart Building/Smart City training, testing and certification lab at the Energy Innovation Center in Pittsburgh, PA [13].

Dickens, Corey; Boynton, Paul; Rhee, Sokwoo.
”Principles for Designed-In Security and Privacy for Smart Cities.”
Paper presented at Cyber-Physical Systems and Internet-of-Things Week (CPS-IoT Week 2019), Montreal, Canada. April 15, 2019 - April 18, 2019.

The core team is a public-private partnership composed of the City of Pittsburgh, a private technology company, and a technology innovation entity.

This project was implemented with no cybersecurity framework, but the private technology company chose and implemented a turnkey plug/Wi-Fi solution with embedded security features.

3.4 Underground Infrastructure Sensing and Mapping for Smart Maintenance, Sustainability, Usage and Resilience, Burlington, Vermont

The Vermont project is a utility-based project for underground infrastructure maintenance and resiliency. This project centers around an autonomous sensor node that can be networked to provide infrastructure monitoring and 3D mapping. The project uses sensing and information technology to determine the state of infrastructure and provide it in an appropriate, timely, and secure format for the users. Information processing techniques convert the data in information-laden databases for use in analytics, graphical presentations, metering and planning [14].

The primary team is a public-private relationship consisting of two cities in Vermont (Burlington and Winooski), a private technology company, and a university.

The Vermont city entities administered overall management and enabled access to city infrastructure, while the private company developed the technical solution. The university researchers addressed security concerns and integrated cybersecurity measures into design decisions, considering the NIST cybersecurity framework.

4 Designed-in Security and Privacy Principles

From the research of the GCTC smart city projects and more extensive examination of a few existing use cases, a set of commonly addressed principles for smart city projects with designed-in security and privacy has emerged: specific technology usage, implementation of a cybersecurity management process, and cybersecurity expertise and public-private partnerships.

4.1 Specific Technology Usage

Most of the projects investigated in this paper clearly focus on a specific technology such as San Leandro's use of moving target data protection technology (Table 2) that is most suitable for a technical use case study rather than a general Internet of Things (IoT) platform. This ensures the security and privacy considerations during planning and implementation of the project can be focused and efficiently managed for specific use cases. Successful examples of specific technology usage can be more readily transferrable and usable by other use cases with similar goals.

Table 2 Specific technologies from example projects

Specific Technologies
Intelligent street light with moving target data protection technology
Wifi smart plug with embedded security features
Infrastructure sensor node considering NIST cybersecurity framework
Traffic monitor with built-in security protocols

4.2 Implementation of a Cybersecurity Management Process and Framework

Many of the projects investigated considered the use of a systematic approach to developing the cybersecurity elements of their smart city/community project, such as San Leandro's efficient street light project. This included following a defined management process (Table 3) and applying a defined framework, such as the NIST Cybersecurity Framework (CSF) used in the Vermont municipalities example, where the framework provides a foundation for understanding and managing security risks (Table 4) [15].

Table 3 Cybersecurity management components

Management
Pre-evaluation
Strategic planning
Operative planning
Implementation
Post-evaluation

Table 4 Core topics of NIST Cybersecurity Framework [15]

NIST Cybersecurity Framework Core Functions
Identify
Protect
Detect
Respond
Recovery

Dickens, Corey; Boynton, Paul; Rhee, Sokwoo.
"Principles for Designed-In Security and Privacy for Smart Cities."
Paper presented at Cyber-Physical Systems and Internet-of-Things Week (CPS-IoT Week 2019), Montreal, Canada. April 15, 2019 - April 18, 2019.

4.3 Cybersecurity Expertise and Public-Private Partnerships

Most of the example use cases relied on cybersecurity expertise from private sector partners to help guide their project. This includes communities who do not have extensive, in-house cybersecurity expertise, such as the City of Pittsburgh and Las Vegas projects, which collaborated with vendors. The inclusion of expert partners from academia or industry to drive cybersecurity design and integration for the project was a major consideration shared by the example use cases. Understanding how such a partnership can work in practice, impacts the productivity of the planning and implementation of security and privacy considerations in these projects.

5 Conclusion

After collaborative discussions with GCTC partners to understand security and privacy aspects of their technology integration and reviewing several projects in the GCTC database, the following common principles and shared best practices were identified for security and privacy consideration and implementation:

(1) Specific Technology Usage;

(2) Implementation of a cybersecurity management process and framework; and

(3) Cybersecurity Expertise and Public-Private Partnerships.

Four example projects were described in this paper to illustrate how these principles and practices can be implemented to develop smart city projects that improve city infrastructure with designed-in security and privacy.

ACKNOWLEDGMENTS

The Department of Homeland Security Science and Technology Directorate partially sponsored the production of this material under a Financial Transaction with the National Institute of Standards and Technology.

DISCLAIMER

REFERENCES

[1] PwC US. *Smart Cities: Five Smart Steps to Cybersecurity.* https://www.pwc.com/us/en/services/consulting/cybersecurity/library/broader-perspectives/smart-cities.html. Published Dec. 2017. Accessed on Nov 9, 2018.

[2] Lohrmann, Dan. *Securing the Smart City. Government Technology*. https://www.govtech.com/security/Securing-the-Samrt-City.html. Published April/May 2018. Accessed on Dec 17, 2018.

[3] Data Security Council of India: Nasscom Initiative. PwC India. *Creating Cyber Secure Smart Cities*. https//www.pwc.in/assets/pdfs/publications/2018/creating-cyber-secure-cities.pdf. Published 2018. Accessed on Dec 17. 2018

[4] NIST. *Framework for Cyber-Physical Systems*. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-201.pdf. Publish 2017. Accessed on Nov 10, 2018.

[5] NIST. Global City Teams Challenge. *Smart and Secure Cities and Communities Challenge*. http://pages.nist.gov/GCTC/. Accessed on Nov 6, 2018.

[6] SmartAmerica Challenge. https://smartamerica.org/. Accessed on Feb 7, 2019

[7] Rhee, S. *Catalyzing the Internet of Things and Smart Cities: Global City Teams Challenge*. 2016; 1st International Workshop on Science of Smart City Operations and Platforms Engineering in partnership with Global City Teams Challenge, Pages 1-4, Vienna, Austria, April 11, 2016, IEEE Explore Digital Library

[8] Global City Teams Challenge Action Cluster. https://gctc.opencommons.org/Category:ActionCluster. Accessed on Jan 31, 2019

[9] GCTC Wiki. *CryptoMove San Leandro Smart Lights Project*. https://gctc.opencommons.org/CryptoMove_San_Leandro_Smart_Lights_Project. Accessed on Feb 4, 2019

[10] ComputerWeekly.com. *Open Information Security Maturity Model (O-ISM3*). https://ww.computerweekly.comehandbook/Open-Information-Security-Maturity-Mode-O-ISM3. Published Jan. 2011. Accessed on February 2, 2019.

[11] Wikipedia contributors. OSSIM. Wikipedia. The Free Encyclopedia. January 7, 2019. https://en.wikipedia.org/w/index.php?titles=OSSIM&oldid=877265361. Accessed on February 2, 2019.

[12] GCTC Wiki. *GENIVI Las Vegas Connected Vehicle Pilot Project*. .https://gctc.opencommons.org/GENIVI_Las_Vegas_Connected_Vehicle_Pilot_Project. Accessed on Feb 4, 2019

[13] GCTC Wiki. *Pittsburgh building portfolio cyber-secure, real-time utility data integration and AI analysis*. https://gctc.opencommons.org/Pittsburgh_building_portfolio_cyber-secure,_real-time_utility_data_integration_and_AI_analysis. Accessed on Feb 4, 2019

[14] GCTC Wiki. *Underground Infrastructure Sensing and Mapping*. https://gctc.opencommons.org/Underground_Infrastructure_Sensing_and_Mapping. Accessed on Feb 4, 2019

[15] NIST. *An Introduction to the Components of the Framework. Cybersecurity Framework*. http://www.nist.gov/cyberframewwork/online-learning/components-framework. Accessed on Nov 6, 2018.

Dickens, Corey; Boynton, Paul; Rhee, Sokwoo.
"Principles for Designed-In Security and Privacy for Smart Cities."
Paper presented at Cyber-Physical Systems and Internet-of-Things Week (CPS-IoT Week 2019), Montreal, Canada. April 15, 2019 - April 18, 2019.

FEA solver integration framework

Jerome Szarazi
(Koneksys, United Kingdom);
Conrad Bock
(US NIST, United States);

Abstract

Integrating finite element analysis (FEA) with systems engineering (SE) would improve traceability, consistency, interoperability and collaboration between SE and FEA activities in multiple engineering disciplines. The first step in achieving this is a software-independent description of FEA models, which are characterized by numerical approximations of partial differential equations (PDEs) derived from physical laws, and finite elements representing unknown physical quantities. In previous work, we presented a finite element mathematics specification that is formal and understandable by most engineers. It provides all information needed for generation of shape functions for physical quantities. In this work, we propose a specification of physics in FEA to complement our earlier mathematics specification.

We first compare existing FEA physics descriptions and their software implementations to highlight the benefits of domain-independent model descriptions used by PDE solvers. A significant drawback of PDE representations is they do not show all physical quantities from which they are derived. To tackle this, we represent the physical laws and derivations needed for FEA PDEs in human- and machine-readable graphs. Instead of classifying physics problems by the kind of PDE, as in PDE solver packages, we formalize problems as paths through these graphs. This increases transparency by capturing modelling decisions currently done on paper or in electronic documents.

We combine the graph-based specification of FEA physics above with the finite element mathematics specification developed earlier to generate linear system of equations (algebraic FEA models) for solving the problem numerically. This combination will enable FEA engineers to design their own libraries (potentially automatically) if they choose, or associate existing solvers. It also generalizes mappings from physics to FEA models, a task currently repeated across specific disciplines. The framework could be standardized and integrated with SE modeling languages, improving interoperability and collaboration between systems and FEA engineers.

Bock, Conrad; Szarazi, Jerome.
"FEA solver integration framework."
Paper presented at NAFEMS World Congress 2019, Quebec City, Canada. June 17, 2019 - June 20, 2019.

1. Introduction and motivation

Simulation-driven design has increased efficiency of product development using computer aided design (CAD) and repeated evaluation of these models with simulations. Software tools for simulation, such as finite element analysis (FEA), computational fluid dynamics (CFD), multibody dynamics (MBD) fall into the category computer aided engineering software (CAE).

Among these kinds of simulations, FEA is one of the most popular numerical methods to test virtual models. For example, FEA helps find the location of maximum stress on a body and therefore assess if and where a product may fail. After identifying weak design locations in simulation results, the design can be digitally corrected using CAD software. By alternating between design and simulation, the virtual model can be refined to eventually meet product requirements. This reduces the need for expensive physical prototypes to test designs. This simulation-design loop helps improve decisions at an early stage of engineering by choosing the best solution candidates in the design space.

Deciding whether a solution meets all requirements also includes reaching agreement between stakeholders. Cross-functional development teams, parallel processes and integrated CAE are necessary to be competitive, but efficient collaboration between FEA engineers, designers and other disciplines is essential for a project success. Difficulties in sharing information and miscommunication increase the time to reach agreement over design alternatives. Such problems are further complicated by use of multiple discipline specific models, often uncoupled and regenerated throughout the design cycle.

Systems engineering (SE) is increasingly used to overcome these problems. Traditional SE provides methodologies, processes and documentation to support such cross-disciplinary development process. In recent years, model-based SE (MBSE) significantly increased efficiency by exchanging models instead of documents [1]. In an MBSE scenario, systems are described by information models, in the same way software code is described by an architecture model.

The Systems Modeling Language (SysML®) [2], an extension of the Unified Modeling Language (UML®) [3], is the common language used to describe and exchange models among system engineers. These system models are digital artefacts, similar to 3D dimensional designs or simulation models, enabling mappings between other simulation data and SysML. Such transformation or integration is facilitated by standard data exchange for simulation models in particular domains. This reduces integration to the interfaces between the domain standards and SysML. For example, Modelica [4], an open and standardized language for time-only simulations, has a standard integration with SysML [5]. Data integration of simulation models with SysML enables traceability to requirements. When information exchange at the interface between a simulation model and SysML is bidirectional, it can lead to simulation automation by synchronizing system requirements changes

Bock, Conrad; Szarazi, Jerome.
"FEA solver integration framework."
Paper presented at NAFEMS World Congress 2019, Quebec City, Canada. June 17, 2019 - June 20, 2019.

with simulation parameters and testing validity of system modifications with simulations. This is also true of simulation optimization, a process critically dependent on bidirectional model mappings.

For FEA, there is, unfortunately no open language to create models or exchange them. As a result, model-based integrations of FEA are platform specific. The absence of standardization leads to point to point integration between FE solvers, often a highly laborious manual or individually scripted process that must be repeated between each pair of points, which is expensive to maintain and sensitive to software updates.

To overcome these problems, we propose a new language for future standardization that helps share reusable information about FEA library models and generate finite element code, with inputs that are understandable by most engineers. This language goes beyond the specific usage of FEA, because it can be applied to physical system modeling in general. It enables FEA engineers to understand relationships between physical quantities within a physical theory and capture modelling decisions. It formalizes the otherwise time consuming modelling process currently done on paper or electronic documentation.

Section 2 describes FEA workflow, data generated at each stage, and reviews standards or formats that support data exchange. Section 3 outlines challenges currently hampering data exchange, one related to converting CAD to mesh data in the context of simulation driven design, and another related to the describing library elements independently of tools. Section 4 presents a new FEA solver integration framework that uses graph structures to capture physical laws, modelling decisions and all model parameters required for a numerical solver. Associated with a mesh standard with virtual topology capabilities, this framework could lead to a new FEA software ecosystem that helps engineers to create and specify their library models. Section 5 concludes the paper.

2. FEA data and standardization

FEA standardization is complex as it requires harmonizing geometric data exchange between CAD and meshes, a unified description of FEA library models and a common format to share simulation results. These three data models are digital artefacts created during FEA simulation, which has three steps: 1) meshing CAD geometry, 2) solving a simulation model, and 3) processing results [6].

a. Meshing

The first step, meshing, requires access to CAD geometry. Such access is facilitated by International Organization for Standardization (ISO) 10303 STandard for the Exchange of Product model data (STEP), a CAD file format supported by all CAD tools [7]. There are many available automatic and semi-automatic meshing algorithms, classification of such algorithms can be found in [8], as well as open source code that generate meshes from STEP files. A

mesh or grid approximates a desired geometry with a collection of discrete shapes in 2D or discrete volumes in 3D used as the computational elements for FEA. Unfortunately, the STEP format for meshes is not widely adopted, the only popular formats being associated with FEA software or meshing tools [9]. In general, mesh information can be stored using a finite element (shape or volume) to node data structure, where each indexed finite element of the mesh refers only to their nodes coordinates or by storing the complete mesh topology, which adds connectivity between the elements that compose mesh shapes or volumes. The first reduces storage space, while the second optimizes data access to volumes, faces and edges for adaptive FEA algorithms such as h-p refinement [10].

b. Solving

The second step of FEA simulation is selection, parametrization and solving of FEA models. There are two groups of FEA solvers, one that abstracts numerical information, and another that abstracts physical domains, which consequently have different inputs. For both models, spatial regions of the mesh, later associated to boundary conditions values, need to be selected and stored.

The first group of FEA solvers provides domain-specific models to solve mechanical (solid or fluid), electrical or thermal problems or a combination of them (multi-physics problems). Each model, called a library element, is associated with a template where material, additional geometric parameters, and initial and boundary conditions values can be set. These models, often identified by proprietary codes, do not provide source code and numerical choices are unknown most of the time. Simulation documentation contextualize the model and guide set-up of the input deck, most often represented as an ASCII file storing mesh, library model reference codes, and boundary and initial conditions. Graphical user interfaces (GUIs) of modern FEA software help in model set-up.

The second group of FEA solvers are partial differential equations (PDE) solvers that uses the finite element method (FEM). PDE solvers, such as Diffpack [11], identify models by PDE type, while others like FENics [12] or FreeFem++ [13] create models by entering the weak or variational form of the PDE, a multilinear functional, constructed by integrating PDEs by parts. The unknown variables of the weak form, usually physical quantity kinds, are substituted by one or more interpolation functions, called finite elements. Domain-independent models are defined with symbolic expressions using a domain-specific language like FreeFem++ or multi-purpose language such as Python [14] for FEnics.

c. Post-processing

The third and final step of FEA simulation is post-processing and evaluation of results. After solving the linear equations, solutions are post-processed if other model quantities are required. For example, in mechanics,

stress fields can be post-processed from displacement fields. Domain-specific solvers store relations between model quantities in their closed source code. Writing additional post-processing code is very often required for domain-independent software.

Evaluating results is facilitated by visualizing physical field quantities by projecting results onto the meshed geometry. Most commercial software solutions provide visualization. The open source visualization platform, Paraview [15], seems to be the most popular open source solution. Its VTK format [16], serialized in a proprietary text format or XML [17], is a de facto standard to exchange simulation results.

3. Current problems and proposals

There are two main problems related to FEA data exchange. First, design-simulation loops are more frequent, but current geometry standards do not facilitate integration between disparate CAD and CAE models, i.e., design geometries and analysis meshes. The other is lack of formal identification of numerical and physics models, which could be resolved using domain-independent models.

a. CAD/CAE integration

The traditional FEA workflow (see subsection 3.b) has five types of digital assets: detailed CAD models or master CAD models, digital mockups,[1] CAD STEP files, FEA mesh proprietary files with boundary conditions, and FEA results files. These are regenerated and maintained whenever the master CAD source is modified. Strategies for CAD/CAE integration framework are detailed in [18].

An intermediary neutral CAD file format, e.g., ISO 10303, provides platform independency between CAD and CAE software, but introduces substantial work due to geometric errors and meta-data losses during the conversion process from CAD proprietary formats to the STEP standard. Geometric errors are due to proprietary modelling of tolerances by CAD systems, requiring correction of the resulting “dirty geometry” [19][20]. In addition, most STEP translators usually omit construction history, parameters and constraints [21], and do not maintain storage of other meta-data, e.g., part names or material data.

The benefit of parametric associative CAD software is that modifications of some parameters, e.g., length, will update all related geometry downstream as well as keep constraints and meta-data associations. Product functionality is maintained by separate administration of geometry and constraints. Ideally, CAE information, such as material and boundary conditions, should be

[1] FEA is usually preceded by either defeaturization or idealization of the detailed CAD design. The resulting simplified models, called digital mock-ups help evaluate system assemblies or kinematics, perform MBD or FEA simulations and visualize FEA results on assemblies or parts [23].

maintained at the CAD level in order to be reused between CAx applications, e.g., MBD and FEA. Most CAD programs provide data interfaces, e.g., product data management (PDM), to enrich CAD models with metadata.

To provide CAD/CAE integration, current trends in CAD technology lean towards solutions that combine CAD and CAE in the same environment, further reducing compatibility between platforms [22]. Before future geometry standards arrive, an intermediary solution could be developing a standard that enables CAD environments to exchange mesh data structures with virtual topology capabilities, as described in [23], by keeping topological correspondence between CAD and mesh. Metadata associated with regions of the mesh could with a standard be exchanged between FEA application and CAD. In comparison to curved geometries in CAD systems, tessellated geometries with polygons faces do not suffer the same problems with tolerances. By moving the CAD/CAE interface from CAD to mesh, workflow efficiency of parametric CAD software is still maintained and platform independence of CAE solvers could be guaranteed. A general topology mesh data structure that captures the connectivity of all mesh elements, as detailed in [25], could maintain topological correspondence between CAD and mesh, enabling mesh regions of interest to be defined, such as material, interfaces or boundary conditions, improving data access, e.g., generation of internal nodes for higher degree finite elements. The mesh could be exchanged with the JSON data format [26], a key-value pairs data structure, that would facilitate data access to the mesh and the virtual topology. The keys, acting as pointers, could link CAD metadata and simulation information to geometric information required by the solver.

b. Formal CAE model identification

Another problem is lack of formal standards that provide enough information to run a simulation with any solver interchangeably. STEP AP209, the standard for FEA model exchange [27], targets mainly domain-specific software, capturing software name, version, and software models identified with proprietary reference codes. The standard provides taxonomies to classify simulation models, but they are too informal to associate a specific solver. Model reconciliation with STEP AP209, evaluating if two models are equivalent, is not possible because source code and numerical assumptions are closed for domain-specific software. This is problematic for solution migration when companies decide to change FEA software; for supplier collaboration when two companies use different software; and for long-term archiving if the software is not supported in the future. Resolving these challenges associated with STEP AP209 is undertaken as a common effort by various groups at different levels. For example, LOTAR's (Long Archiving and Retrieval) Engineering Analysis and Simulation Workgroup (EAS) develops, publishes and maintains standards for archiving and retrieval of key FEA input/output characteristics at various stage of development in a robust and repeatable fashion [28]. For collaboration, MOSSEC (Modeling and Simulation

information in a collaborative Systems Engineering Context), develops methods for organizing and sharing Modeling and Simulation meta- data and information in a collaborative system, and for capturing context to enable traceability [29]. Tools to analyze and visualize STEP file are developed by NIST [30].

In comparison, domain-independent software solutions require details about numerical choices. Their models are reusable in multiple physical domains that have the same mathematical structure. For example, a physics problem described by a Laplace PDE can be used to solve thermal conduction as well as electrostatics problems. Symbolic code is very often used for model definition, which can be automatically compiled into a lower level C code. Symbolic code was in use in the 90s for finite difference in the Sinapse framework [31] and in the later 90s for finite elements [32].

The weak form equation, formalized in [33] as a language and extended in [34] as a unified framework for finite element assembly, complemented by a finite element and user defined function, is enough information to symbolically or numerically integrate each cell of a mesh and assemble a linear system of equations that is interpretable by a linear algebra solver (LAS). Solving the systems of equations at this stage only requires linear algebra methods, such as Cholesky or Krylov subspace methods that many frameworks support, such as Petsc [35] with parallel processing capabilities or libraries compatible to the BLAS specification [36].

However, writing symbolic expressions for PDE weak forms is uncommon for FEA engineers. Even though PDEs are widely-understood and useful for model classification, using PDEs for model identification is problematic because boundary conditions refer to variables not in the PDEs. Without documentation detailing a PDE's derivation from multiple equations, they provide only a partial view of the model, limiting post-processing of other model variables. Such post-processing requires expert knowledge of model relationships to derive finite elements of post-processed variables. In contrast, domain-specific software models hide finite element choices and variable relationships, but offer simpler post-processing capabilities.

Ideally FEA model definition should be understandable by most engineers and generate solutions with any solver. Starting with weak forms, we can reverse-engineer the workflow used to find their expressions. Weak forms, can be derived from PDEs using integration by parts, enabling weak forms to be linked to their corresponding PDEs or automatically derived. Next PDEs are derived from model equations that correspond to physical laws. Transitioning from domain-independent descriptions, the mathematical model, to domain-dependent description, the physics model, is done by substituting mathematical variables with physical quantities. The resulting physics model is understandable by engineers and can be used by numerical methods such as finite volumes (FV) or finite differences (FD). For FEA models, finite elements that describe unknown physical quantities and space-dependent parameters are needed, in combination with physics models formulated as PDE weak forms to

construct linear algebraic systems of equations that can be solved by LASs.

To achieve formal description of FEA library models, we need a formal description of finite elements as well as an efficient mechanism to describe physics equations and the corresponding derivation of PDEs. The rest of the paper discusses these two aspects in more detail.

4. New FEA solver integration Framework

In this section, we present a new platform-independent approach to integrate FEA solvers. To provide an open model specification, subsection 4.a introduces building blocks of a graph-based language to capture physical equations, as a starting point for model definition. Linking these equation graphs together captures the mathematical structure of physics describing the system being analyzed. Subsection 4.b explains how math or physics problems are defined by selecting known and unknown variables in math and physics graphs. This captures modeling decisions as functional programs, enabling extraction of all problems as math/physics evaluation subgraphs or math/physics solver subgraphs. Subsection 4.c details how boundary condition types can be automatically collected and associated to math/physics solver subgraphs. To facilitate reusability, subsection 4.d introduces the concept of common mathematical structure that can be reused for multiple domains, illustrated in subsection 4.h with the rapid extension of library elements. Once problems are defined, subsection 4.e explains how physical quantities are represented in FEA by finite elements, with examples on how to use our finite element specification. Subsection 4.f clarifies the association of finite elements to physics solver subgraphs by creating numerical solver graphs and generating input templates for solvers. In subsection 4.g, an example presents how a mesh standard with virtual topology capabilities would associate design and FEA metadata and, together with the library element specification of subsection 4.e, provide a platform-independent description of FEA solvers.

a. Equation models

Whereas time-only (a.k.a., lumped parameter, 1D, network) simulation handles simple topologies (e.g., electric circuits) of many different library elements (e.g., capacitor, resistor…), FEA simulates one or few library element(s) on space embedded topologies (meshes). FEA library elements are more complex than time-only library elements and space is multi-dimensional and multi-directional in contrast to time, which flows in one direction.

Time-only solutions have open standards, such as Modelica or proprietary languages, such as XCOS [37] and Simscape/Simulink [38], to describe libraries or models, which are defined with physics equations. For signal-based simulations, mathematical expressions of Laplace transformations are either captured in custom or standard library blocks that can be connected together.

In contrast, FEA solutions deliver libraries as complied code that are parametrized with input templates. This transparency problem can only be solved with an ecosystem shift that gives FEA engineers the same flexibility as

time-only simulations. To achieve this, we believe physics equations should also be inputs to define FEA libraries. This would help in modeling FEA libraries, which is currently done on paper or electronic documents.

Mathematical structures of space-dependent physics models (physical equations describing a domain), are captured by Tonti diagrams [39] and used in [40] to illustrate FEA variational principles. Tonti diagrams are formalized with the cell method, a discretization method based on algebraic topology, first introduced by Branin [41], that defines topological dualities for both space and time as well as cell-specific mesh discretizations.

FEA engineers are not familiar with the cell method, so we propose a more general approach, one that first captures equations as triples, two physical quantities linked by an operator. Triples are linked whenever they share the same elements; a graph is automatically created. These graphs define symbolic computation, as compared to Simulink graphs that define real number processing. With this approach, structures described by Tonti diagrams can be captured in a way more accessible to engineers.

To explain how this works, let's consider an engineer who aims to build an FEA library with this approach. His first step would be to find relevant physical equations in the literature. These could be captured using a GUI that helps select and connect necessary blocks, or by entering symbolic equations that are progressively displayed as graphs. Figure 1 shows the gradient law for relating potentials and gradients.

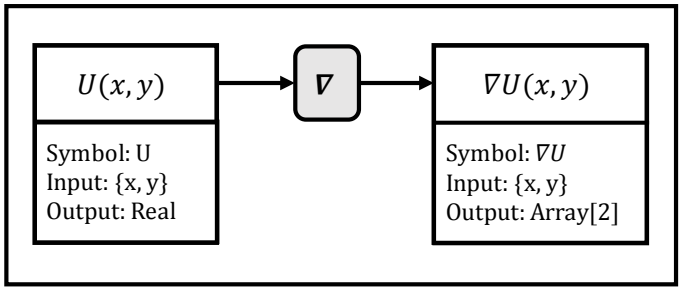


Figure 1: Description of the gradient law

Math objects in physics are tensor fields or tensors. Tensor fields are functions, in physics, they map from space coordinates to a real number or arrays of real numbers. In the first case, tensor fields are scalar functions while in the second they are arrays of functions. Each math object is defined with respect to a Euclidian space of dimension N, the space on which library elements are defined, and symbols for each dimension. (e.g., 2D space with $S = \{x, y\}$). We can restrict the space of a map by choosing a subset of the coordinate symbols. Math objects that have no input coordinates are tensors, which is a real number or an array of real numbers that only depend on the coordinate system but not the coordinates. Figure 2 shows type declarations (symbol, input, output) of some math objects and their representation. It

defines math objects for a two-dimensional space with $\{x, y\}$ as symbols for real number coordinates. For example the math object with symbol U is a scalar field that takes $\{x, y\}$ as input, and outputs a real number. Another example is the math object K, a constant, which is coordinate-independent (no input) and outputs a real number.

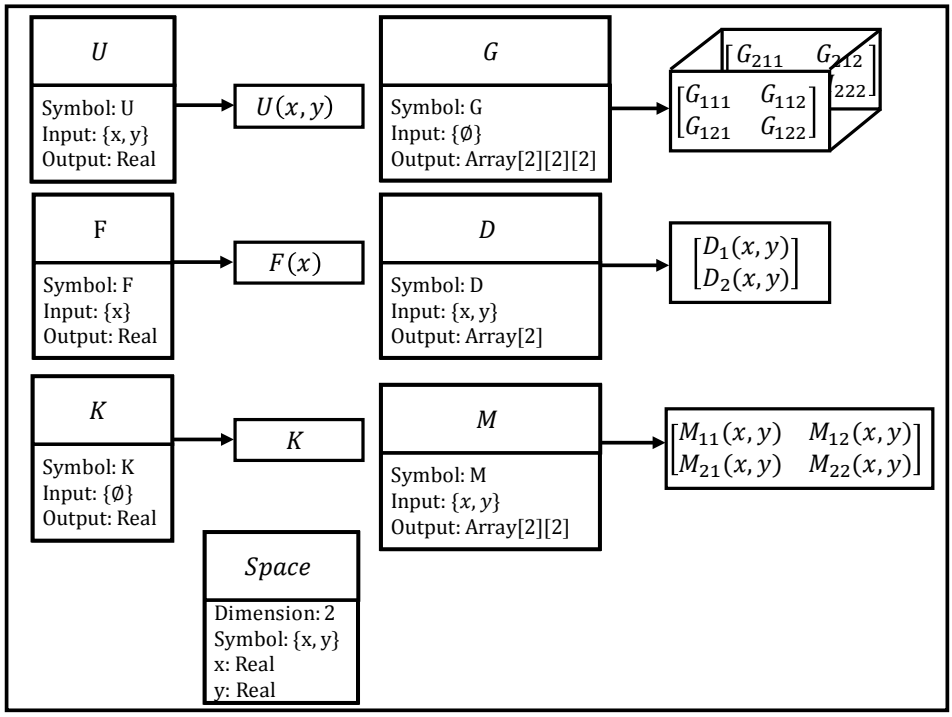


Figure 2: Declaration of math objects

Math objects are linked in a *math graph* (MG), e.g., one for the gradient law as shown in Figure 1. MGs are bi-partite and directed, which by definition are composed of two sets, one for math objects and another for operations linking two math objects. Each edge indicates how math objects relate to the operators at its ends. An arrow coming into an operator comes from a math object input to the operator, while an arrow going out of an operator leads to a math object output from the operator. The operator transforms one math object into another. Having operator as nodes in MGs enables representation of binary operators such as addition and multiplication. To support unambiguous references, each node is unique (identified, e.g., by a unique resource identifier on the web) and to support specialization, each has multiple attributes. Math objects can be augmented (*specialized*) with units and symbols following ISO [42] to produce *physics objects*. In Figure 3, the engineer entered the physical law that defines temperature gradient. For model completeness, the inputs of each math object have been added, which are the space or time coordinates or a combination of these.

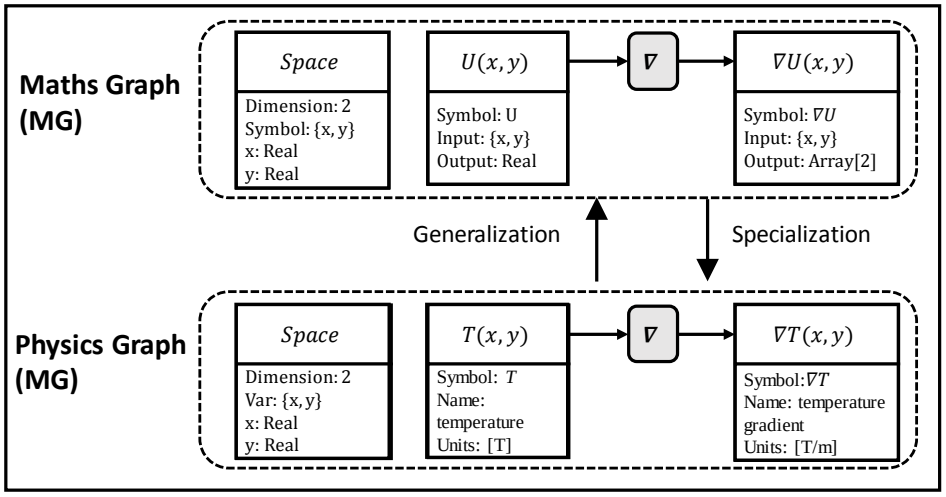


Figure 3: Specializing a physics graph (PG) from a math graph (MG)

As the number of relevant equations increases, they can be automatically combined when two equations, each a statement or triple linking two physical quantities by an operation, share the same physical quantity, as illustrated in Figure 4. Combination is possible only if each physical node is uniquely identified. For example, the Resource Description Framework [42], the web standard for linked data, could identify them with unique resource identifiers (URIs) and describe equations with two statements each, one linking a physical quantity to an operator, and another linking that operator to another physical quantity.

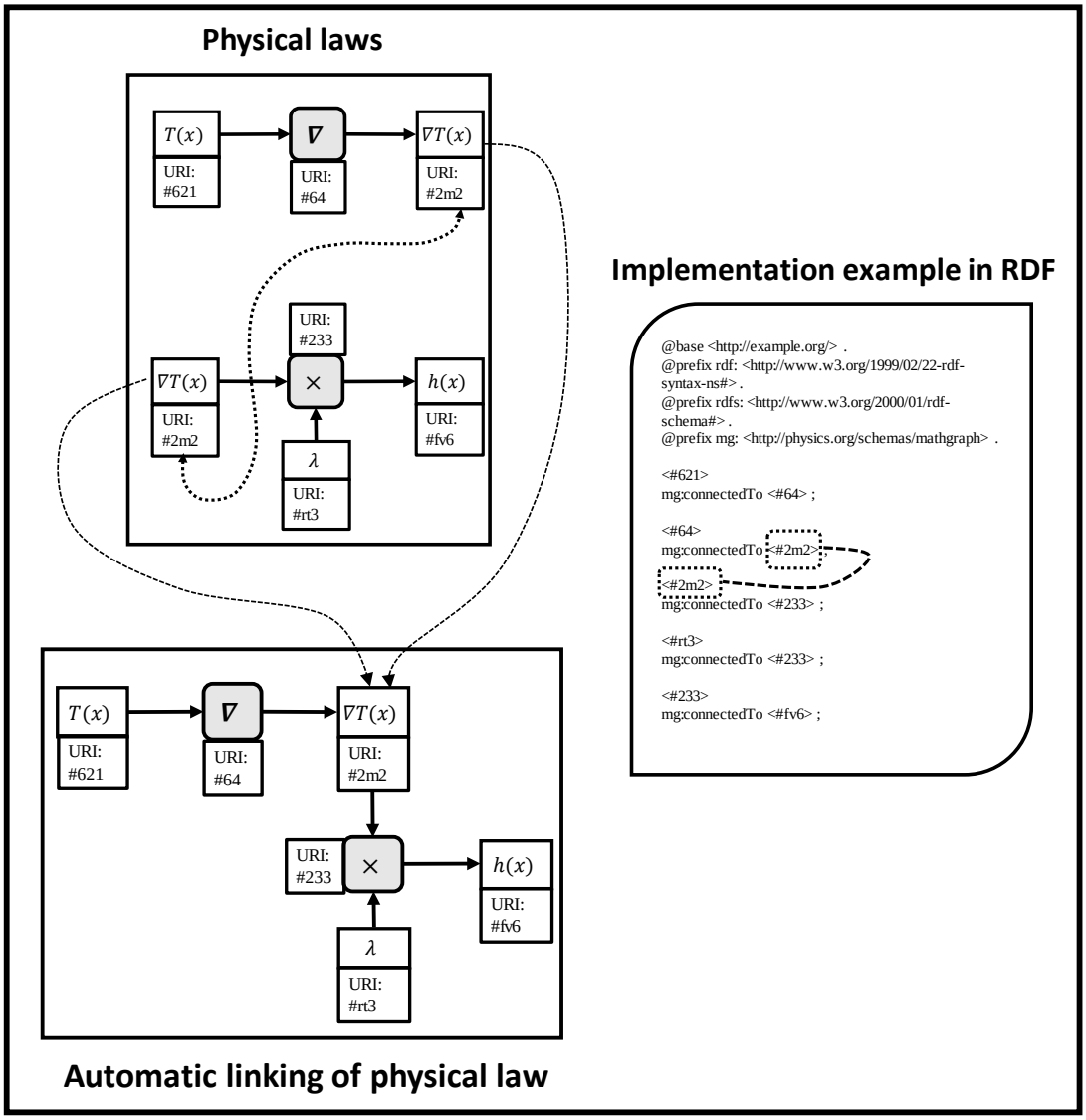


Figure 4: Automatic linking of physical equations

Combining equations, automatically or manually, produces a *physical graph* (PG), a combination of physical equations that capture the mathematical structure of a physics theory or model. We distinguish three layers of a PG as shown in Figure 5. The *functional program specification layer* defines the type of math or physics objects and operators, and the objects input/output for each operator (arrows). A path through a graph in the direction of arrows from one object to another is called a *functional program path (FPP)*. Objects between the start and end of a path are intermediary results. The second layer is the *expression layer* that declares a mathematical expression (e.g., e^{5x^2}) specifying

outputs in terms of inputs. The types of math and physics objects can be checked for consistency with the expression (e.g., gradient of scalar field produces a vector field). This facilitates expression processing when mathematical expressions, e.g., LaTeX [44], MathML [45] or expressions trees, are assigned to math or physics objects. Symbolic computation can produce the expression for a math object from the expression of another and the operation linking them, in the direction of the link. The *data layer* specifies input values mapped to output values following the rule defined by the expression, creating a function graph or plot.

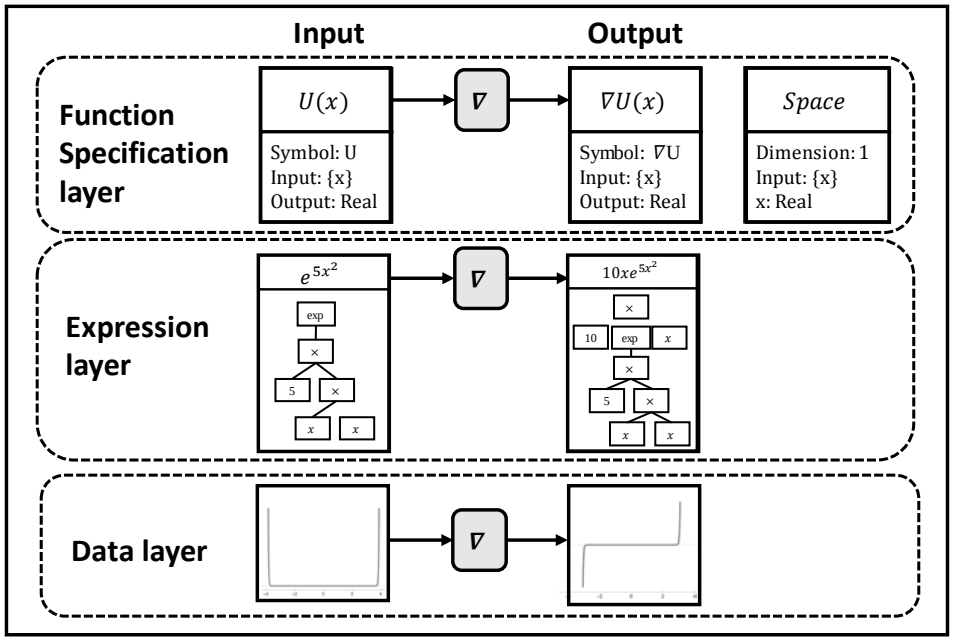


Figure 5: Layers of a math objects

b. Physics problems

After defining math and physics models, or choosing existing ones, the next step specifies problems to solve by declaring known and unknown pairs of math or physics objects in MGs and PGs. If a mathematical relation exists for a pair, then at least one FPP exists through the MG or PG with the two math or physics objects at the ends of the path. Absence of path means no solution is possible without modifying the graph. If a FPP has operations composed only of invertable functions, then an *inverse* FPP can be defined by replacing the operations by their inverses and reversing the arrows in the path.

When known and unknown are assigned as the start and end of a FPP, respectively, we call the path a *math evaluation graph (MEG) or physics evaluation graph (PEG).* When known and unknown are assigned as the end and start of a FPP, respectively, and an inverse FPP exists, an MEG or PEG can be automatically generated for it as shown in the Figure 6.

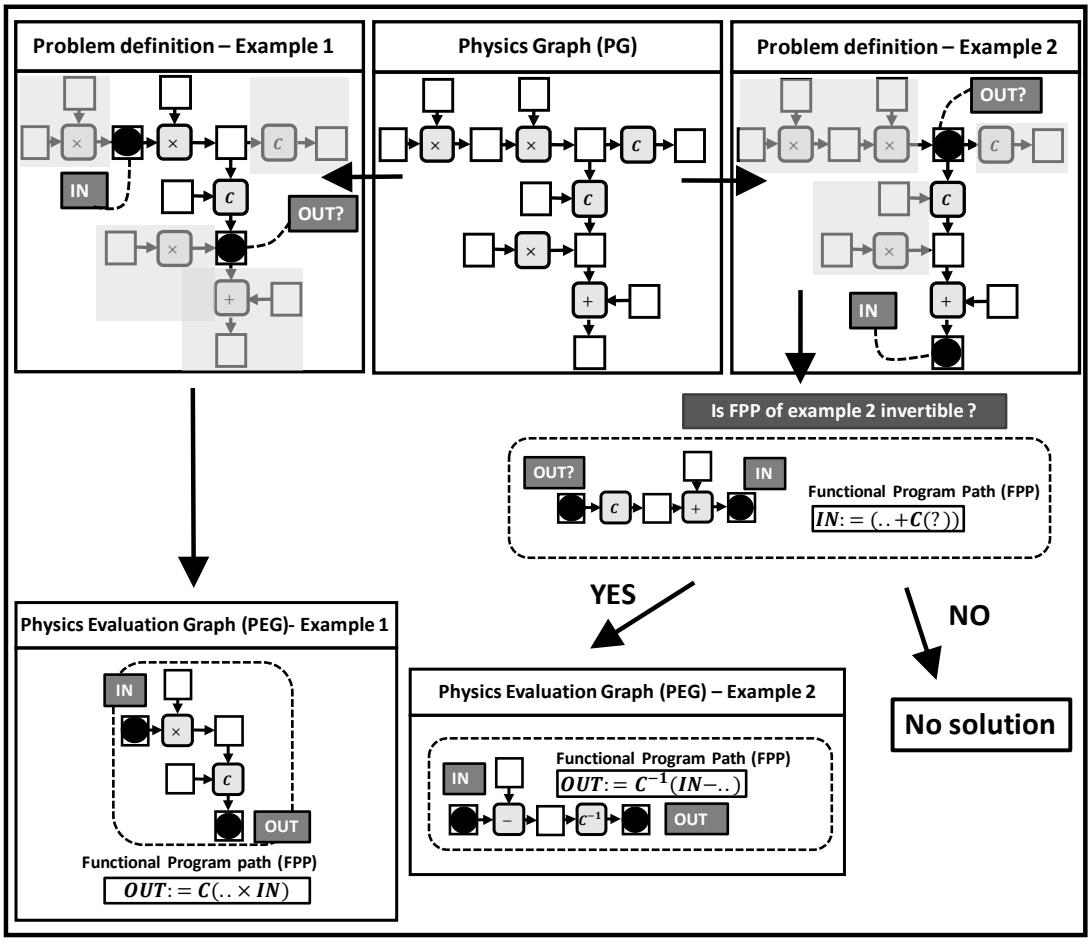


Figure 6: Solution finding in a physics graph

In physics, most laws are expressed in mathematical equations using differential operators, which can be inverted by adding boundary conditions. For example, in one dimension (e.g., motion on a line, see Figure 7), the top PEG connects a position/force pair, while the middle one inverts it by adding integration constants as boundary conditions for the anti-derivative operations. For example, the time derivative of position is velocity, on the left in the top PEG, but its inverse, integration, in the middle PEG, produces distance, rather than position. A specific position value, a boundary condition, is needed to get position from distance. Similarly, on the right, the time derivative of momentum is force, but the integral of force is impulse. A specific boundary momentum is added to get momentum from impulse.

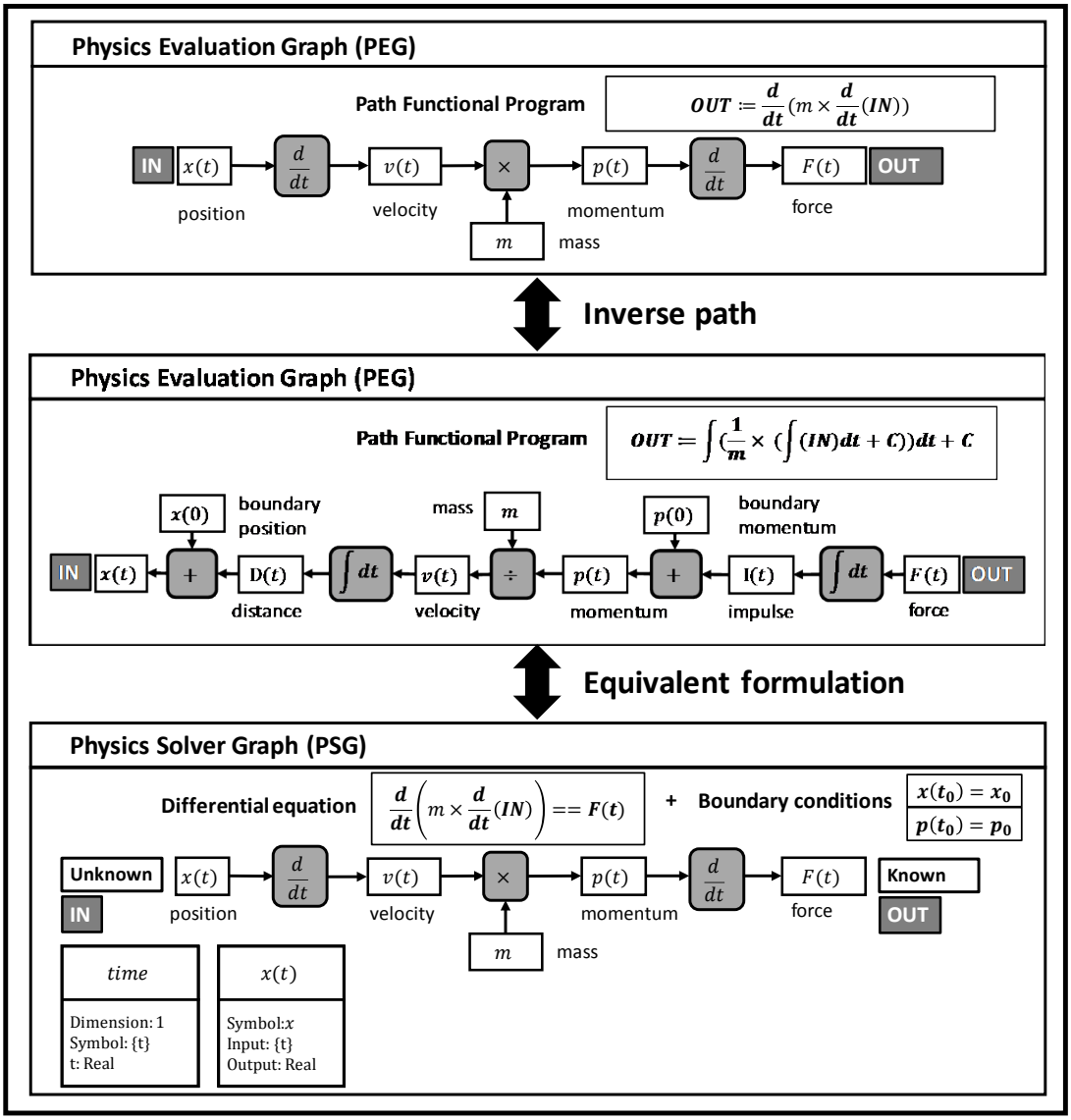


Figure 7: Inverse path and its equivalent formulation for ODEs

A more common representation of such problems is differential equations, which in one dimension, ordinary differential equations (ODEs), can be handled by symbolic solvers. When a PEG has integration operators and a known start object (e.g., force known, position unknown in the middle graph of Figure 7), its inverse PEG has differential operations and an unknown start object (e.g. position). Because the end object (e.g., force) is known and must be equal to the differential expression implied by the rest of the graph, the graph represents a differential equation. To be equivalent to the original integral PEG, we must add boundary conditions (e.g., the lower graph of Figure 7).

This is a *math solver graph (MSG)* or *physics solver graph (PSG)*. Paths through these graphs go from unknown to known under boundary conditions, representing differential equations under boundary conditions.

Most physical laws are expressed in differential form are represented as a differential operator linking two physical quantities. For multi-dimensional spaces, there is no analytic inverse to differential operators, therefore, analytic inverses of most FPPs with differential operators do not exist. The problem of finding the inverse can still be characterized by an MSG or PSG, see Figure 8. An approximate solution to the inverse FPP can be constructed either by using an equivalent formulation to the PDE (e.g. FEM) or directly discretizing the operators (e.g. FD or FV).

In our context, most graphs will have differential operators. The methodology consists of defining known and unknown object pairs, then finding a path connecting these objects. Most of the time there will be only one possible path. We can generate a MEG or PEG when the start object is known, or a MSG or PSG if the start object is unknown. The two graphs are the same except for the known/unknown choice for the start and end objects, and additional boundary conditions attached to the MSG or PSG, which will be detailed in subsection 4.c. MEGs and PEGs are used for evaluation problems or post-processing, while MSGs and PSGs are used to specify solver problems that involve ODEs, PDEs or systems of PDEs.

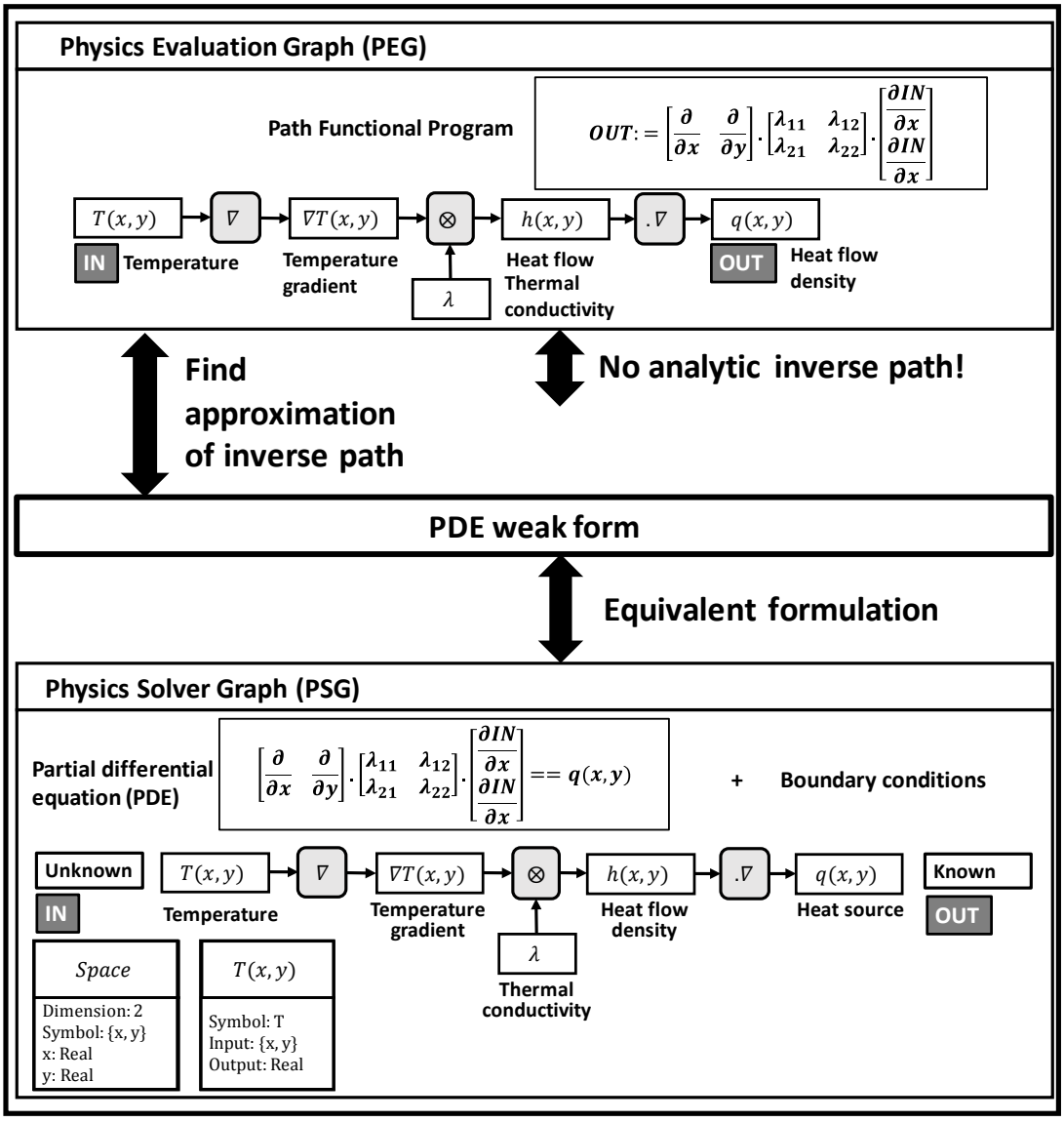


Figure 8: Inverse path and its equivalent formulation for PDEs

Characterizing or finding PDEs is useful to identify a problem, because problems are classified by PDEs, but they are difficult to understand without their derivation from physical laws. Furthermore, their boundary conditions are expressed in relation to the unknown, the start node of FPP. PDEs only represent the mathematical relationship between start and end objects in a FPP. Physical quantities and operators along the path are not in the PDE. In contrast, MSGs or PSGs give the derivation of PDEs. The graphs are more than equations, they capture modelling decisions leading to equations.

c. Boundary conditions

Finding a unique numerical PDE solution requires boundary conditions (BCs), along with other parameters (material, geometry). Specifying BCs is the task of mathematicians. Some BCs are well-known such as Dirichlet, Neumann, Robin, and Cauchy BCs. Mathematicians prove existence of solutions to PDEs under specified constraints. For solver input completeness, it would be very useful to automatically associate BCs during construction of PDEs or ODEs, but this would require solution existence proofs to be automated, which is not currently possible.

However, characterizing the physical quantity associated to BCs is still useful to engineering. We call this quantity, the *BC value type* (e.g., electric potential). PSGs give the derivation of PDEs and ODEs helping determine which physical quantities can be post-processed but requires BCs. Only adding their BC value types is possible. BC value types are associated directly to differential operators. Following Stoke's theorem,

$$\int_{\partial\Omega} w = \int_{\Omega} dw$$

each differential operator applied to a math object and integrated over a region has a corresponding boundary value type for the result of the integral of the math object along the boundary of the region. If we consider the triple (input math object, operator, output math object), then the BC type is linked to the input by the boundary integral, as shown in the bottom row of Figure 9. For curl and divergence, this introduces a third physical quantity. For gradient or differential, no additional physical quantity is introduced, the BC type math object is the input to the operator (integral of a single point value is equal to itself).

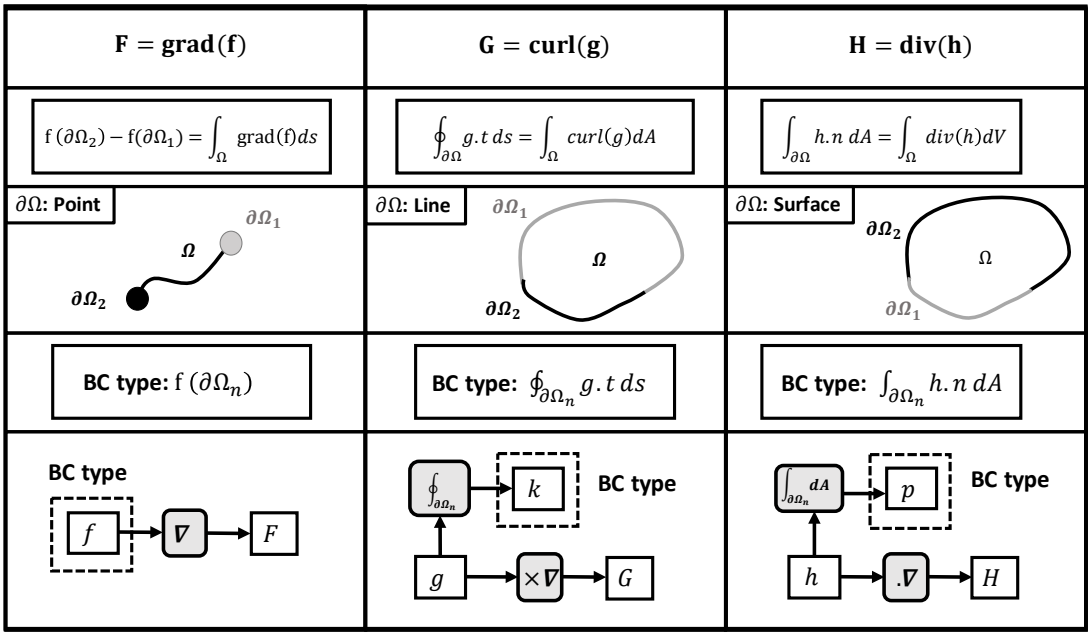


Figure 9: Stoke's law and corresponding BC types attached to differential operators

These BC types will be automatically added in the MG or PG whenever a differential operator is defined. They are important because they also correspond to physical quantities. For example, Figure 10 shows a thermal conduction PG, where heat flow is the BC type of the divergence operation applied to heat flow, and temperature is the BC type of the gradient operator. If we consider the PSG with temperature unknown and heat source as known, temperature and heat flow are the collected BC types which correspond to the Dirichlet BCs type and Neumann BCs type respectively. By automatically adding BC types, whenever a differential operator is used, we have a PG complete model.

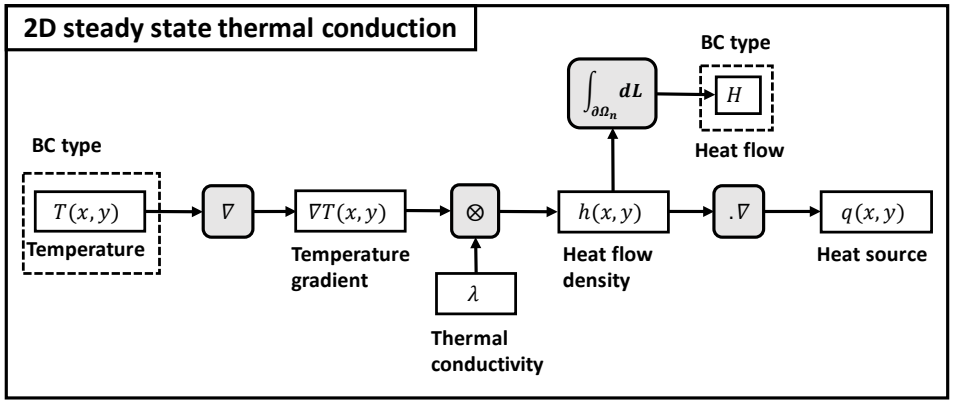


Figure 10: 2D steady state thermal conduction PG

d. Common mathematical structure

A *common mathematical structure* (CMS) is an MG that is shared by many PGs. Identifying CMSs is useful for reuse (e.g., assessing solver compatibility) and model understanding (e.g., from a mathematical perspective solving thermal conduction problems is the same as solving electrostatics problems). For example, FEA solves classical field theory problems. The CMS in Figure 11 is the basis of many domain-specific problems (e.g. 2D steady thermal conduction in Figure 10). Benefits of CMS will be illustrated in subsection 4.h.

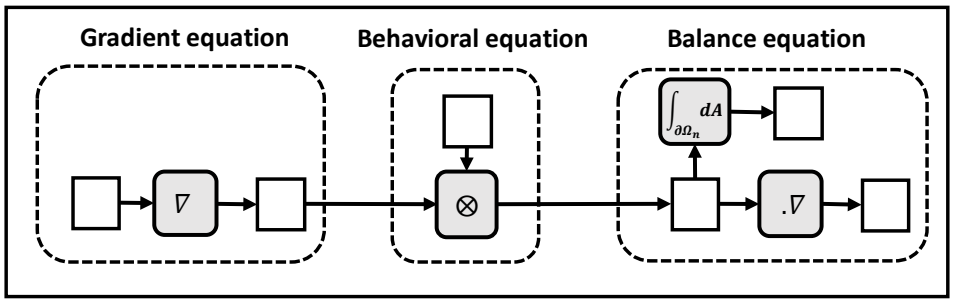


Figure 11: An example of a CMS of classical physics

e. Finite element specifications

As described in section 4.b, problems defined by an MSG don't have an inverse but an equivalent description the weak form. The discretization process consists in defining the problem in the integral form in order to find an approximation of the inverse path of the FPP. Finite elements (FE) are functions or tensor fields representing physical quantities on discrete topological elements of space. Degrees of freedom (DOFs) of finite elements are free or fixed variables that represent evaluations of physical quantities done on topological elements (point, line, surface, volume) . Functions or tensor fields of FEs, usually represented by polynomial (or tensor) functions, are rearranged as shape functions or shape fields, each fully defined, and each one multiplied by one DOF coefficient. There are as many shape functions as DOFs. For example, in electrostatics, on the left in Figure 12, a potential can be interpolated by a FE with 3 DOFs, specifically 3 electric potentials at discrete points in space as free variables. Another example, from thermal conduction, would be heat flow density interpolation, on the right in Figure 12. This time, normal components of heat flow density are integrated along each edge, leading to 4 DOFs or 4 heat flow as free variables.

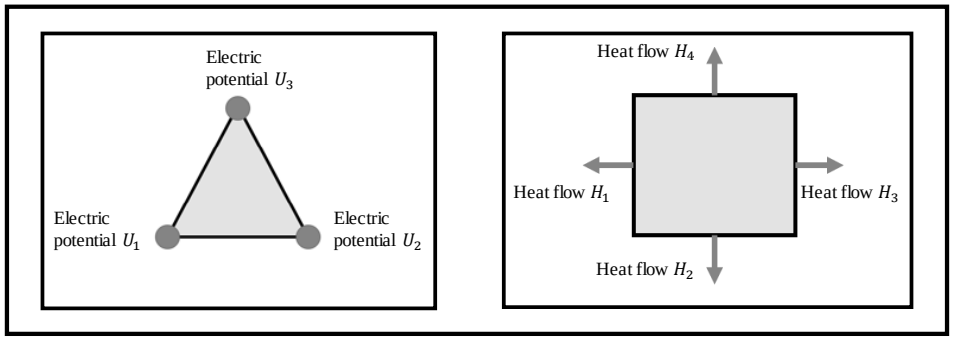


Figure 12: DOFs for electric potential and heat flow density interpolation using finite elements

Finding a formal FE description is complicated by the same finite elements being referred to as many different names in the literature. For example, a linear line element is also called a Lagrange line element. Engineering names, such as beam or bar element, are also ambiguous. A beam, for example, can have 4 degrees of freedom or 2 degrees of freedom., A finite element periodic table in [46] classifies these, but requires advanced mathematical knowledge to understand.

We developed a formal finite element specification (FES) [47] based on the generic finite element definition of Ciarlet [48] and the DOF type description as found in [49]. However, the triplet (geometry, DOF and basis space) is described with topology to provide compact description. A central aspect of

FES is associating DOFs to topological elements (point, line, surface, volume). A geometry Ω, like a triangle for example, can be partioned into a set of three lines $\mathbf{C}_1(\Omega) = \{\mathbf{L}_1\ \mathbf{L}_2, \mathbf{L}_3\}$ corresponding to disjoint surface boundaries; a set of three points $\mathbf{C}_0(\Omega) = \{\mathbf{P}_1\ \mathbf{P}_2, \mathbf{P}_3\}$, the edge boundaries. Subtracting the $\{\mathbf{C}_0(\Omega), \mathbf{C}_1(\Omega)\}$ union from Ω gives $\mathbf{C}_2(\Omega)$, the interior surface. Each member of $\mathbf{C}_2(\Omega)$, $\mathbf{C}_1(\Omega)$, $\mathbf{C}_0(\Omega)$ are sets of points; surface, edge, singleton (one member) point sets respectively. DOFs are defined by associating DOF types, e.g., point evaluation (PE), first derivative (FD), along with a natural number to $\mathbf{C}_0(\Omega)$, $\mathbf{C}_1(\Omega)$ and $\mathbf{C}_2(\Omega)$. For example $\{PE: 1\}$ associated to $\mathbf{C}_0(\Omega)$, written $\mathrm{D}_{\mathbf{C}_0(\Omega)} = \{PE: 1\}$, means point evaluation for each member of $\mathbf{C}_0(\Omega)$, therefore 3 and 4 PEs (DOFs) for triangles and squares respectively. $\mathrm{D}_{\mathbf{C}_1(\Omega)} = \{PE: 1\}$ means one PE at midpoint for each line point set of $\mathbf{C}_1(\Omega)$. In the triangle case, on the right in Figure 13, the evaluations are at 3 midpoints, one for each line. In general, for PE on lines, points divide lines into regular partitions, e.g., a midpoint divides a line into two equal parts. Multiple DOF types and number can be assigned to each $\mathrm{D}_{\mathbf{C}_n(\Omega)}$. A FE is explicity specified by composing all necessary $\mathrm{D}_{\mathbf{C}_n(\Omega)}$.

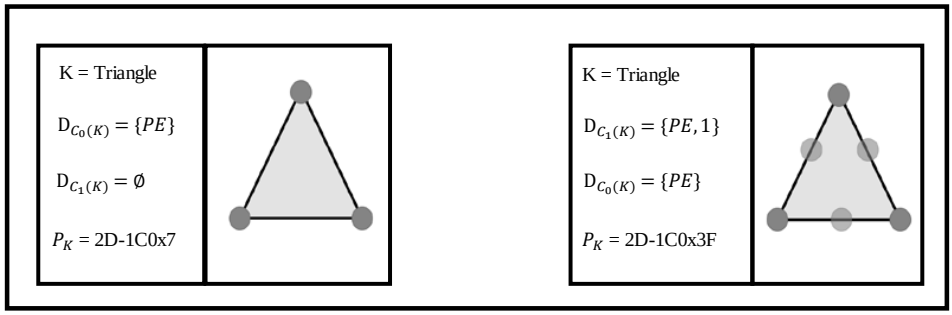


Figure 13: Example of FE definition using the FES

With geometry and DOF type specification of an FE, it is possible to calculate the number of DOFs and find an appropriate function to represent all DOFs. The FES can also encode polynomial functions. The set of all possible monomials can be lexicographically ordered, so the set of monomials appearing in a given polynomial can be mapped to a bit sequence, where each bit corresponds to the presence of a particular monomial in the polynomial. The resulting sequence is encoded as a hexadecimal number. The decoding of the hexadecimal number only requires writing the monomials for the space dimension in a graded lexicographic order, then assigning to each monomial position a bit that clarifies whether the monomial is used or not. For example, Figure 14 shows the encoding of a two-dimensional scalar polynomial (e.g. 2D-1C), more specifically a quadratic serenpedity functional space, functional space defined on a square, that uses all monomial terms except x^3, y^3 for the set of monomial $\{1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3\}$. The main idea of the encoding is to provide an implementation description that is independent of

domain-specific ontologies (e.g., serenpedity). The current encoding method is on possibility as an example how domain-independent descriptions could work.

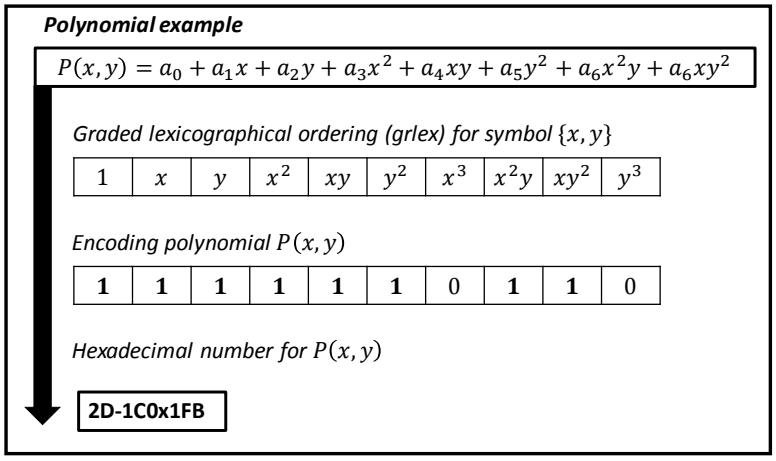


Figure 14: Example of functional space encoding

f. Library element

After creating PGs and PSGs, FEs are specified and associated to physics objects. FEs represent unknowns and space-dependent parameters (material or geometric parameters) if necessary. To illustrate the process, we consider the PG of a bar element shown in Figure 15, created using the kinematic equation that links strain to displacement; Hooke's law, which links stress to strain; and the balance equation that links body load to stress. Force is added automatically as a result of the divergence operator used in the balance equation. Then a problem is specified with a PSG, e.g., body load-displacement as known-unknown pair and displacements and forces as BC types. This is one of 12 evaluation and solver problems that can be defined using the same PG of Figure 15.

Next, the association of FEs to physics objects of the PSG, defines numerical graphs (NGs), by first declaring the math type of the physics object and then the FE. Many NGs can be defined from one PSG. For the body load/ displacement pair example, displacements could either be linear or quadratic FEs. Constant, linear, quadratic FEs to describe space-dependent sections are also possible, which in total, with displacement options, would produce 6 NGs. At this point, math objects consistency can be checked, verifying whether FEs choices (e.g., linear or quadratic) are consistent, helping understand numerical choices. In Figure 15, if distributed load is constant, choosing a linear displacement FE would be inconsistent, because balance equations would be broken and we would have only an approximation.

The principle of virtual displacement, which only requires kinematic admissible displacements, allows such approximations. However, inconsistent

physical equations are frequent when mixed principles introduce additional unknows in a PSG. In Figure 15, choosing quadratic displacement with constant body load yields an exact solution. These examples show that many numerical models can be defined using the same PSG, which itself is one of many options derived from a PG. This is reflected in the large number of library elements in software packages.

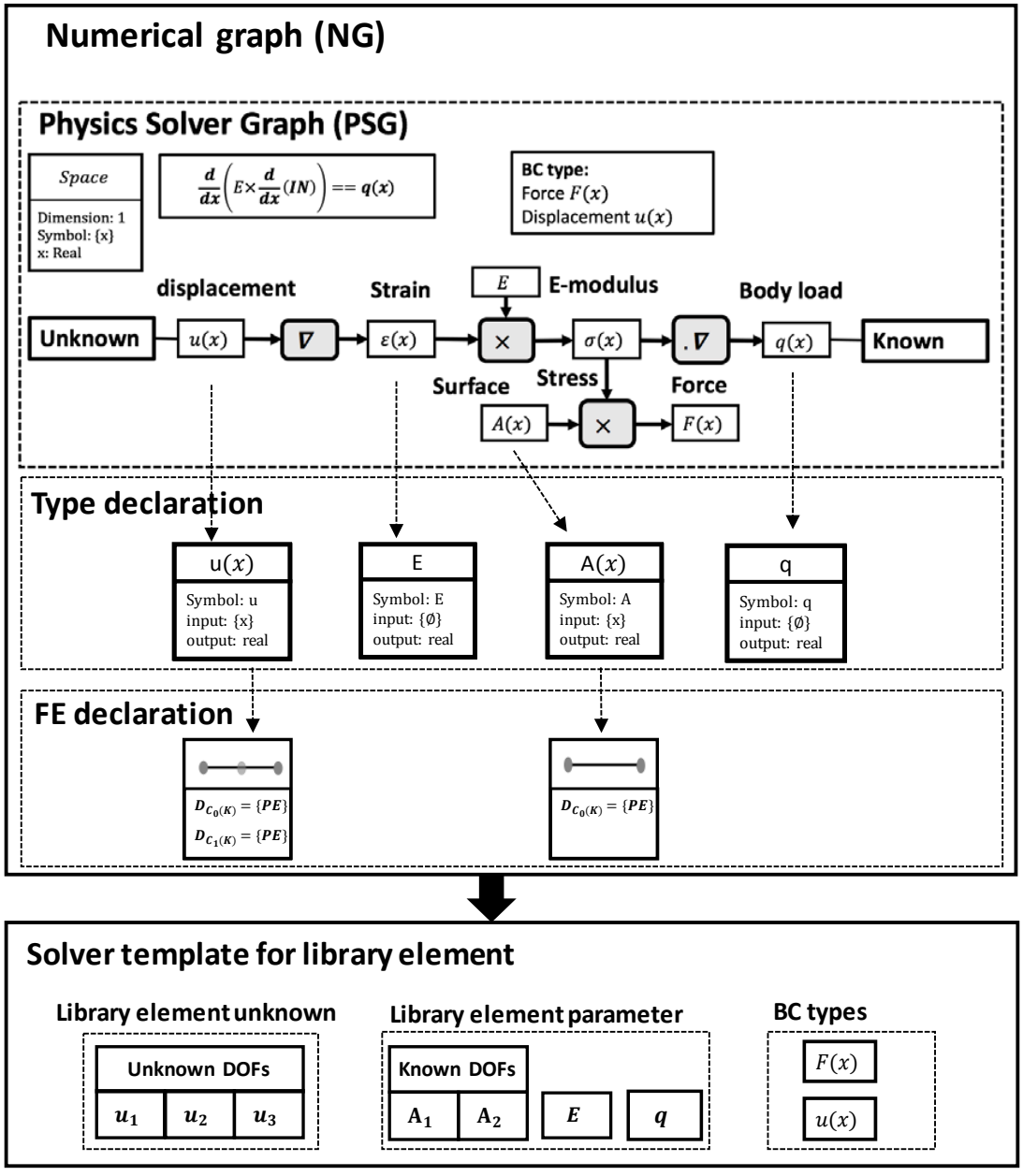


Figure 15: Example of a Numerical Graph for a bar

Specification of library elements leads to at most four kinds of digital artefacts: mathematical models as MGs (math graphs), physics models as PGs (physics graphs), physics problems as PSGs (physics solver graphs), and a subset of PGs and numerical models as NGs (numerical graphs).

NGs identify solvers (PDEs), collect all solver inputs (BC types, FEs, physical quantities with units and relations), information to construct the PDE weak forms, an integral formulation of the PDE, required for assembling FEA simulations. In addition, NGs, being specialization of PSGs, enable post-processing once the solution is found.

g. Simulation

In simulation driven design, traceability between design and simulation is essential, yet often obscured conventional practice. As an example of description granularity and information traceability, rather than optimize simulation, consider a part that is attached to another plate with two screws, as illustrated in Figure 16. A force is applied at the unattached end. The part is created using CAD with plate length as parametric variables. If meshing would natively occur in CAD environments instead of exporting CAD files to pre-processors or CAE tools, if the resulting mesh could be exchanged using a mesh standard that is serialized with popular file formats (e.g., JSON) and defined with schemas capable of representing BREP (boundary representation) mesh topology that associate metadata to topological elements, as described earlier, then CAD meta information would be preserved and reusable by any solver capable of reading such mesh standard. The mesh would store topological elements (node, edge, faces, volumes), the topology (how elements connect) and regions of interest (virtual topologies). CAD uses BREP and can associate design metadata to bounded shapes (e.g. material, assembly information), which can be translated with a native mesh process to virtual mesh topologies that keep metadata associations. The mesh could be exported with two separate information sets, one that describes the mesh topology and region topologies with keys, the other that associates keys with meta-data (e.g., material). In our example, the mesh has 4 virtual topologies; the plate (with material information and basis geometry for the mesh), the two holes with assembly information (use of screws) and a section of the boundary region (force). A new mesh with consistent metadata can be created when length or other parameters are updated in CAD.

The next step is selecting a library element, which requires FEA expertise and library knowledge. Sometimes a package for the required simulation is not available (e.g., license or absence of the library element, probably not the case for this problem). Searching for library package solutions requires reading documentation. Comparing these models is difficult because documentation standards vary and packages are not open enough. A standard based on the graph data structure previously presented would simplify the library search process. In the example above, search would be, e.g., for NGs with triangle

shapes containing physical objects such as force and displacement. In this example, starting with a PG of 2D elastostatics, the problem can be identified (here we know that force and displacement are the BC types), therefore a PSG connecting displacement to body load is required to have these two BC types. Then, a NG is created by associating a finite element to physical quantities or space-dependent parameters if necessary (e.g. space-dependent plate width).

The next step is to create a simulation model. In the example above, the NGs are associated to the plate domain and NG BC types (e.g., displacement and force) are associated to mesh regions (e.g., displacement to boundary holes). Once the simulation model is complete, solvers that follow this standard could propose their services. Another option is a new ecosystem of software that could support library creation and automatic generation of library code or links to existing code.

The final step is simulation. Each simulation run is defined by simulation parameters (material data, e.g., shear and bulk modulus, derived, e.g., from steel information, thickness value).

As a result of this approach we have 5 artefacts, design metadata, mesh data, NG (library model), simulation model metadata (attaching BC type to mesh data), and simulation data (referencing simulation value to NG parameter). In addition, the neutral mesh links design metadata to simulation model meta. A change in the CAD geometry (e.g., side length) does not impact the simulation model data. Furthermore, if an FEA engineer changes the FE geometry specification, for example changing triangular to square mesh, the CAD software could read this information and update the mesh accordingly .

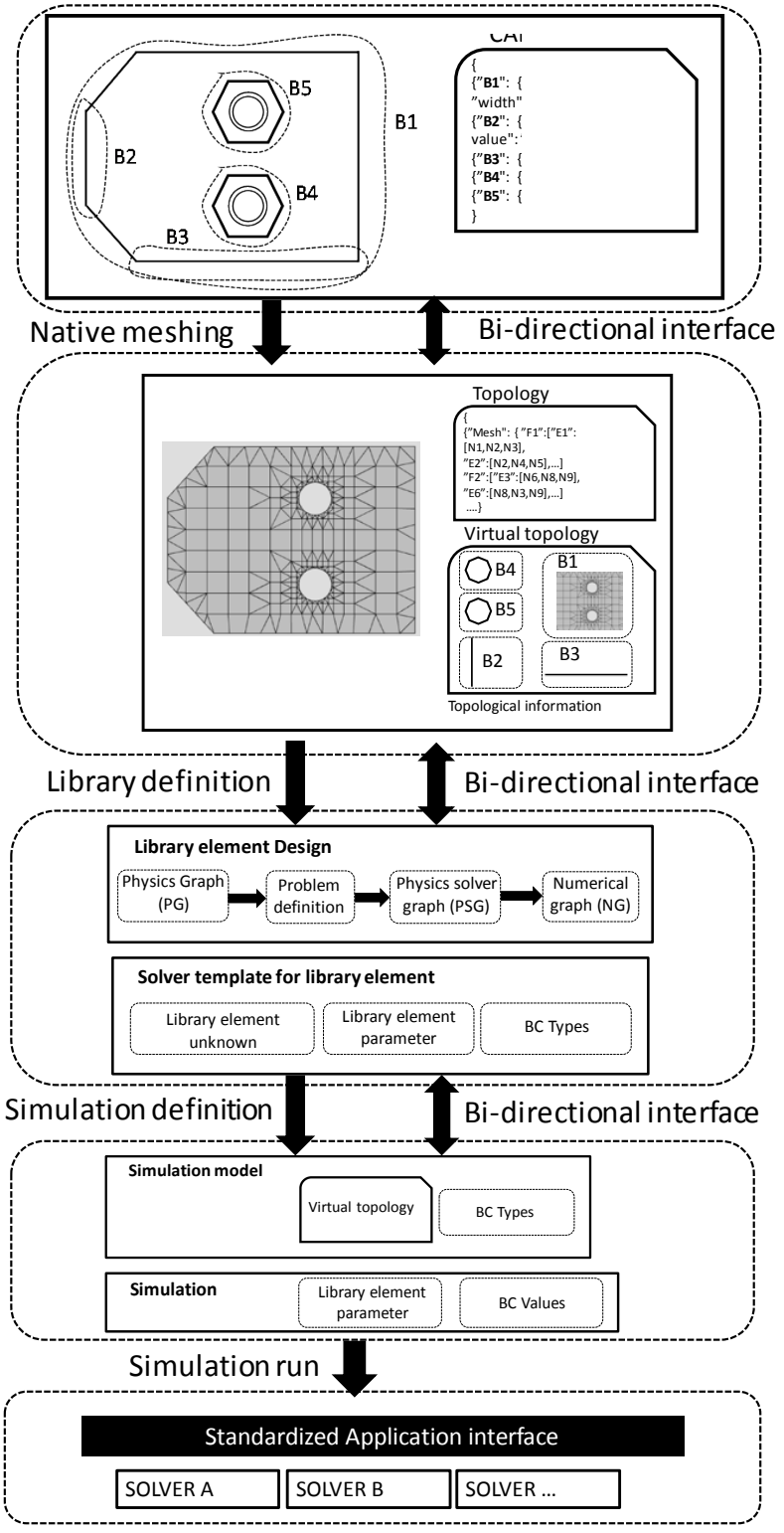
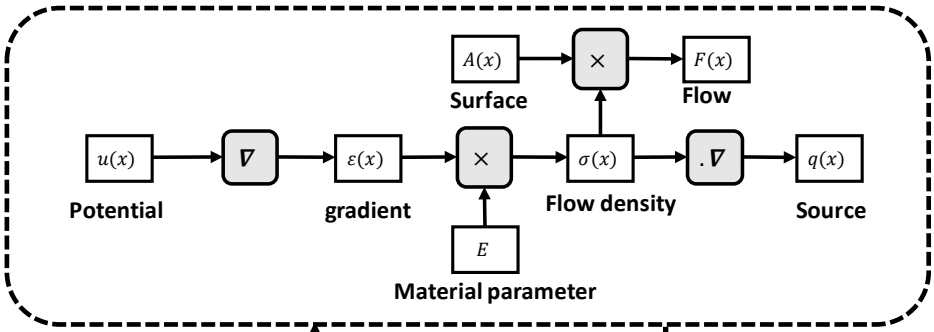


Figure 16: Standardized integration of solvers with assumption of a mesh and FEA library standard

h. Rapid extension of libraries

Identifying CMSs helps with reuse because of the specialization process that produces PGs from MGs, PSGs from PGs and finally NGs from PSGs. For example, a single MG handles most one-dimensional space-dependent PGs used for FEA. Such core models are essential to rapidly extend libraries. From this one MG, 10 PGs can be derived using correspondence tables. MG node objects are specialized to PG node objects by adding symbols and units. Physics objects and math objects with same mathematical relations in a graph are linked by specialization. This enables existing libraries to be rapidly extended to other physical domains when models share the same MS. Figure 17 shows three specialization examples (axial stress, themal conduction, electrostatics, see rows of table) from the 1D space-dependent core model.



Model name	Potential			Gradient			Flow density			Source			Flow			Material parameter		
	name	symbol	unit	name	symbol	unit	name	symbol	unit	name	symbol	unit	name	symbol	unit	name	symbol	unit
Axial Stress	Displacement	$\boldsymbol{u}$	[m]	Strain	$\boldsymbol{\epsilon}$	[-]	Normal stress	$\boldsymbol{\sigma}$	[N/m2]	Body load	$\boldsymbol{q}$	[N/m3]	Force	$\boldsymbol{F}$	[N]	Elasticity modulus	$\boldsymbol{E}$	[N/m2]
Thermal conduction	Temperature	$\boldsymbol{T}$	[K]	Temperature gradient	$\boldsymbol{\nabla T}$	[K/m]	Heat flow density	$\boldsymbol{h}$	[W/m2]	Heat source	$\boldsymbol{\sigma_t}$	[W/m3]	Heat flow	$\boldsymbol{H}$	[W]	Thermal conductivity	$\boldsymbol{\lambda}$	[W/(m.K)]
Electrostatics	Electric potential	$\boldsymbol{U}$	[V]	Electric field	$\boldsymbol{E}$	[V/m]	Electric displacement	$\boldsymbol{D}$	[C/m2]	Charge density	$\boldsymbol{\rho}$	[C/m3]	[-]	$\boldsymbol{\psi}$	[C]	Permittivity	$\boldsymbol{\epsilon}$	[F/m]

Figure 17: Rapid creation of libraries by CMS specialization

Many strategies are possible for library extension. A domain-independent strategy would be to create any possible NGs derived from MGs and MPGs, then specialize them to the corresponding physics, generating many PGs. Another would be to progressively create PGs, PSGs, NGs and identify PGs that share the same MG, then create models that can be reused for other domains.

In addition of being reusable, PGs can be composed with coupling equations, which can either connect PGs of different physics domains, leading in this case to multi-physics PGs; or connect PGs of the same domain, leading

to PGs describing more complex behaviors. For example, a thermal conduction PG and an axial loading PG, both sharing the same MG, can be connected by the thermal expansion equation. This coupling creates so-called thermo-mechanical models.

5. Conclusion

Virtual product development and simulation driven design decreases development costs by reducing expensive physical prototype testing. Early virtual design testing using CAE simulation, such as FEA, helps verify whether designs satisfy stakeholder requirements. Tracing digital models to requirements and system architecture is efficiently managed with model-based engineering, a methodology that facilitates integration and exchange of digital models between engineering disciplines. Standards are enablers of such model exchange.

Unfortunately, current trends show CAD and FEA packages are increasingly natively integrated in the same proprietary environment to leverage benefits of parametric CAD design [22]. This introduces two standardization challenges, one is related to translation errors between CAD formats (meta-data losses and geometric tolerance errors), the other due to informal FEA model description by standards. The first introduces substantial work when transitioning between CAD environments and between CAD and CAE, because of metadata losses. As an intermediate solution before CAx standards solve these issues, the development of a mesh standard that neutrally references topological regions could link PDM/CAD data to FEA data. The second is mainly due to the FEA software ecosystem, which delivers element libraries as compiled code packages. As a result, current standards only identify models with PLM data, such as software name, proprietary references to the compiled code and informal ontologies to characterize models, supporting only point to point integration.

An ecosystem shift is needed to solve the second challenge above, one that gives the same flexibility as time-only simulations and more transparency. Instead of setting inputs of compiled code templates, the starting point to model or generate FEA libraries should be an abstraction level understandable by all engineers, i.e., physics equations. At the numerical abstraction level, FEA engineers detail their FE choices. If both abstraction levels are connected, then FEA engineers and other engineering disciplines share a common understanding of their models. We show how these two abstraction levels connect by progressively creating and storing information as graph datastructures (MGs, PGs, PSGs, NGs). With numerical graphs (NGs), input template for solvers can be generated (PDE, BC types, material and geometry, FE choices). Solvers could be integrated into platforms though standardized

interfaces by providing a formal library element specification, or in the future new kinds of software solutions that model library elements and generate code.

Identifying common mathematical structure, core models, is essential for rapid extension and reuse of libraries between physical domains. In addition, physics graphs can be composed. We see two progressive usage scenarios, one that formalizes the description of FEA libraries (integration), the second that directly links physics models and numerical decisions to solvers with a standardized interface to run simulation with any solvers or even generate code (interoperability and transparency).

6. Acknowledgements

The authors thank Benjamin Urick, Stephen Langer, and Joseph Draper for their detailed comments and helpful discussion.

This work was performed under grant awards 70NANB16H174 and 70NANB18H192 from the U.S. National Institute of Standards and Technology. Identification of any commercial equipment and materials is only to adequately specify certain procedures. It is not intended to imply recommendation or endorsement by the U.S. National Institute of Standards and Technology, nor does it imply that the materials or equipment are necessarily the best available for the purpose.

7. References

[1] BKCASE Editorial Board. (2017). The Guide to the Systems Engineering Body of Knowledge (SEBoK), v. 1.9.1 R.J. Cloutier (Editor in Chief). Hoboken, NJ: The Trustees of the Stevens Institute of Technology. Accessed 2019. www.sebokwiki.org.

[2] Object Management Group (September 2015). OMG Systems Modeling Language*TM*, version 1.4: http://www.omg.org/spec/SysML/1.4.

[3] Object Management Group (March 2015). OMG Unified Modeling LanguageTM, version 2.5: http://www.omg.org/spec/UML/2.5.

[4] Modelica Association (2017) Modelica Language Specification, version 3.4

[5] Dadfarnia, M., Bock, C., Barbau, R. (2016). An Improved Method of Physical Interaction and Signal Flow Modeling for Systems Engineering: Conference on Systems Engineering Research.

[6] Hardwick, M, Clay, R., Boggs, P., Walsh, E., Larzelere, A., Altshuler, A. (2005), “DART System Analysis,” Sandia National Laboratory Report SAND2005-4647, Aug.

[7] International Organization for standardization (2016) ISO 10303-21:2016 Industrial automation systems and integration -- Product data representation and exchange -- Part 1: Overview and fundamental principles

[8] Ho-Le, K (1988). Finite element mesh generation methods: a review and classification, Butterworth & Co
[9] Geuzaine, Christophe & Remacle Jean-Francois (1997-2019). Gmsh – A three dimensional finite element mesh generator with build-in pre- and post-processing facilities – open source www.gmsh.info
[10] Beal, M. W. and Shephard, M. S. (1997), A general topology-based data structure. Int. J. Numer. Meth. Eng., 40: 1573-1596.
[11] Bruaset, Are Magnus & Langtangen, Hans Petter (1998). Diffpack: A software Environment for Rapid Prototyping of PDE Solvers
[12] Alnaes, M.S. & Blechta, J. & Hake, J. & Johansson, A. & Kehlet, B. & Logg, C., Richardson J.& Ring, J.& Rognes, M. E.& Well, G.N. (2015) The FEniCS Project Version 1.5: Archive of Numerical Software. vol. 3.
[13] Hecht, F. (2012). New development in FreeFem++. Journal of numerical mathematics, 20(3-4), 251-266.
[14] Van Rossum, G. (2007). Python programming language: http://www.python.org.
[15] A. Henderson, ParaView Guide, A Parallel Visualization Application. Kitware Inc., 2007.
[16] Schroeder, Will & Martin, Ken & Lorensen, Bill (2006), The Visualization Toolkit (4th ed.), Kitware, ISBN 978-1-930934-19-1
[17] W3C (September 2006) Extensible Markup Language (XML) 1.1 (Second Edition)
[18] Lee S.H., (2005). A CAD-CAE integration approach using feature-based multi-resolution and multi-abstraction modelling techniques: Computer-aided design, Elsevier.
[19] Beall, Mark & Walsh, Joe & Shephard, Mark. (2003). Accessing CAD Geometry for Mesh Generation: Proceedings of 12th International Meshing Roundtable. 33-42.
[20] Wassermann, Benjamin & Kollmannsberger, Stefan & Yin, Shuohui & Kudela, László & Rank, Ernst. (2018). Integrating CAD and Numerical Analysis: 'Dirty Geometry' handling using the Finite Cell Method.
[21] Junwahn, Kim & Pratt, Michael J. & Iyer, Raj & Sriram, Ram (2007). Data Exchange of Parametric CAD Models using ISO 10303-108, NISTIR 7433, NIST
[22] Hirz, Mario & Rossbacher, Patrick & Gulanova, Jana. (2017). Future trends in CAD – from the perspective of automotive industry. Computer-Aided Design and Applications. 14. 1-8. 10.1080/16864360.2017.1287675.

[23] Tierney, C. M., Sun, L., Robinson, T. T., & Armstrong, C. G. (2015). Generating analysis topology using virtual topology operators. Procedia Engineering, 124, 226-238
[24] Hirz, M. & Dietrch, W. & Gfrerrer, A. & Lang, J. (2013), Integrated Computer-Aided Design in Automotive development, Development processes, Geometric Fundamentals, Methods of CAD, Knowledge-Based Engineering Data Management. Springer
[25] Beal, M. W. and Shephard, M. S. (1997), A general topology-based data structure. Int. J. Numer. Meth. Eng., 40: 1573-1596.
[26] W3C (December 2017) The JSON Data Interchange Syntax: http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf
[27] International Organization for Standardization (2014) 10303-209:2014: Industrial automation systems and integration -- Product data representation and exchange -- Part 209: Application protocol: Multidisciplinary analysis and design.
[28] Long Archiving and Retrieval Engineering Analysis and Simulation Workgroup (2019) http://www.lotar-international.org/lotar-workgroups/engineering-analysis-simulation/
[29] MOSSEC (2019) - Modelling and Simulation information in a collaborative Systems Engineering Context - http://www.mossec.org
[30] Lipman, Robert (2018) – STEP File Analyzer User's Guide (Version 5) NIST Advanced Manufacturing Series 200-6. https://doi.org/10.6028/NIST.AMS.200-6
[31] L. Akers, Robert & Baffes, Paul & Kant, Elaine & Randall, Curtis & Steinberg, Stanly & L. Young, Robert. (1998). Automatic synthesis of numerical codes for solving partial differential equations. Mathematics and Computers in Simulation. 45. 3-22. 10.1016/S0378-4754(97)00082-7.
[32] Korelc, Joze. (1997). Automatic Generation of Finite-Element Code by Simultaneous Optimization of Expressions. Theoretical Computer Science. 187. 231-248.
[33] Alnæs, Martin & Logg, Anders & Ølgaard, Kristian & Rognes, Marie E. & Wells, Garth N. (2014) Unified Form Language: A domain-specific language for weak formulation of partial differential equations. Volume 40 issue 2, February 2014, Article No. .ACM Transactions on Mathematical Software (TOMS)
[34] Alnæs, Martin & Logg, Anders & Mardal, K-. A. & Skavhaug, O. & Langtangen, H.P. (2012). Unified Framework for Finite Element Assembly

[35] Balay, Satish & Abhyankar, Shrirang, & F. Adams, Mark & Brown, Jed & Brune, Peter & Buschelman, Kris & Dalcin, Lisandro & Eijkhout, Victor & Gropp , William~D. & Kaushik, Dinesh & Knepley, Matthew~G.& McInnes, Lois Curfman and Rupp, Karl and Smith, Barry~F. & Zampini, Stefano and Zhang, Hong, (2015). PETS, Users Manual, Argonne National Laboratory
[36] Van de Geijn, Robert & Goto Kazushige (2011). BLAS (Basic Linear Algebra Subprograms). Encyclopedia of Parallel Computing.
[37] Scilab (2019) XCOS, open source, https://www.scilab.org/about/scilab-open-source-software
[38] The MathWorks (2019), Simulink/Simscape Documentation, https://www.mathworks.com/help/simulink/, https://www.mathworks.com/help/physmod/simscape/.
[39] Tonti, Enzo (2013) The mathematical structure of classical and relatistic physics: Birkhauser
[40] Felippa, Carlos. (1994). A survey of parametrized variational principles and applications to computational mechanics. Computer Methods in Applied Mechanics and Engineering.
[41] Franklin H. Branin. The algebraic-topological basis for network analogies and the vector calculus. In Proceedings of the Symposium on Generalized Networks, volume 16, pages 453 – 491, Brooklyn, New York, 1966. Polytechnic Institute of Brooklyn
[42] ISO 80000-1:2009 (2009), Quantities and units, ISO
[43] W3C (2014) Resource Description Framework (RDF) 1.1 https://www.w3.org/RDF/
[44] LaTex project (2019) LaTex – a document preparation system - https://www.latex-project.org/about/
[45] W3C (2014) Mathematical Markup Language (MathML) version 3.0 2nd edition https://www.w3.org/TR/MathML3/
[46] Logg, A. & Arnold, D. (2014). Periodic table of finite elements: Siam News.
[47] Szarazi, Jerome & Bock, Conrad (2017) Integrating Finite Element Analysis with Systems Engineering Models, NAFEMS world conference, NAFEMS.
[48] Ciarlet, P. (2002). The finite element method for elliptic problems: SIAM.
[49] Logg A. & Al. (2012). Automated solution of differential equation by the finite element method. Springer.

Proceedings of the ASME 2019
International Design Engineering Technical Conferences
and Computers and Information in Engineering Conference

IDETC/CIE2019
August 18-21, 2019, Anaheim, CA, USA

IDETC2019-98415

A REVIEW OF MACHINE LEARNING APPLICATIONS IN ADDITIVE MANUFACTURING

Sayyeda Saadia Razvi[1], Shaw Feng[1], Anantha Narayanan[2], Yung-Tsun Tina Lee[1], Paul Witherell[1]

[1]National Institute of Standards and Technology (NIST), Gaithersburg, MD, United States
[2]University of Southern California, Los Angeles, CA, United States

ABSTRACT

Variability in product quality continues to pose a major barrier to the widespread application of additive manufacturing (AM) processes in production environment. Towards addressing this barrier, monitoring AM processes and measuring AM materials and parts has become increasingly commonplace, and increasingly precise, making a new wave of AM-related data available. This newfound data provides a valuable resource for gaining new insight to AM processes and decision making. Machine Learning (ML) provides an avenue to gain this insight by 1) learning fundamental knowledge about AM processes and 2) identifying predictive and actionable recommendations to optimize part quality and process design. This report presents a literature review of ML applications in AM. The review identifies areas in the AM lifecycle, including design, process plan, build, post process, and test and validation, that have been researched using ML. Furthermore, this report discusses the benefits of ML for AM, as well as existing hurdles currently limiting applications.

Keywords: additive manufacturing, machine learning, deep learning, data analytics, algorithm, survey, review

1. INTRODUCTION

Additive Manufacturing (AM) is an advancing and increasingly popular manufacturing technology that embodies the revolutionary progress of the modern manufacturing industry [1]. It is a process in which a part is made by joining material, layer by layer, directly from 3D model data [2]. AM offers competitive advantage over traditional manufacturing techniques by enabling fabrication of low volume, customized products with complex geometries and material properties, in a cost-effective and time-efficient way [3]. The rapid proliferation of AM technologies has resulted in seven well-defined sub-categories of AM, several of which are capable of producing metallic parts [2]. With continuing technological advances, AM has evolved from being limited to fabricating prototypes to producing end-use metallic parts in various applications (e.g., aerospace, defense, biomedical, and automotive) [4].

Despite the growth of and advancements in the AM industry, achieving consistency with part quality and process reliability in AM remains a challenge [3]. The fundamental reason for this situation is that both the shape and material properties of a part are formed during the AM process. Realizing any AM part involves intricate design, material, and process interactions over the course of a complex multi-stage process that includes five major steps: designing, process planning, building, post-processing, and testing and validation [5]. The controlled and precise execution of each of these steps is needed to fabricate a qualified part.

Recent efforts to reduce AM part variability have focused on learning as much as possible about parts and processes through monitoring and inspection [6]. Advancements in sensor technologies, sensor fusion and data acquisition methods [7], have led to an unprecedented increase in AM data, encompassing many of the aspects of "big data" (Table 1). The different types of data generated throughout the design-to-product transformation cycle are creating new opportunities (Figure 1) for knowledge discovery throughout AM processes [8].

Table 1. Characteristics of AM Data

Volume	~0.5 TB of in-situ monitoring data per build [9] ~TBs of CT scan data
Velocity	Up to 600 variables logged per second during the build 75 GB/s of image data [11]
Variety	Numerical (machine logs, process parameters) 2D images (thermal, optical) 3D (CAD models, CT scans) Audio (acoustic signals) and videos (thermal, optical)

Razvi, Saadia; Feng, Shaw; Narayanan, Anantha Narayanan; Lee, Yung-Tsun; Witherell, Paul.
"A Review Of Machine Learning Applications In Additive Manufacturing."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE2019), Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

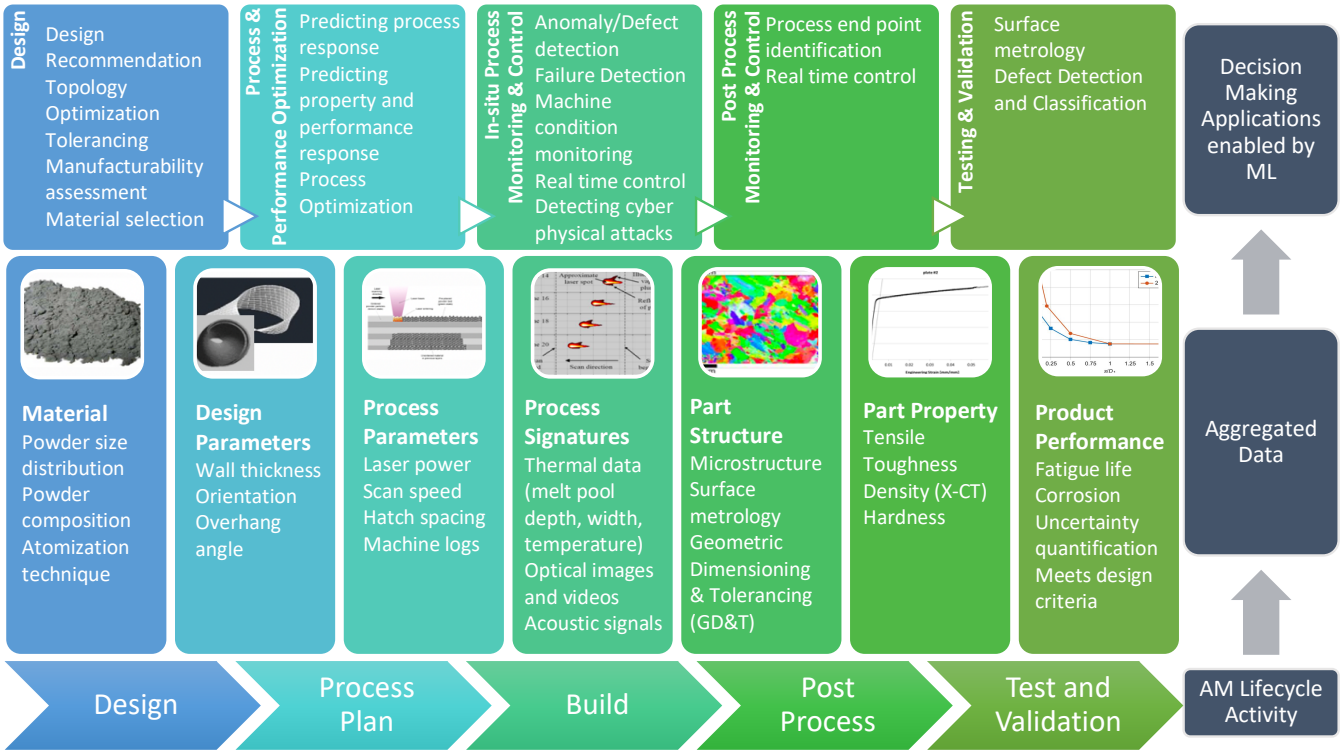


Figure 1. AM Lifecycle, Examples of Associated Data, and Decision Making Applications

In a sense, AM has become a manufacturing domain that is data-rich but knowledge-sparse. Extracting knowledge from the vast amounts of available AM data can be a tedious process. Despite the advances in measurement science and increasing number of datasets from the AM lifecycle, there is limited scientific understanding to characterize AM materials-geometry-process-structure-property-performance relationships. Advanced computational and analytical tools are needed to process the high dimensional and complex data. To this end, new developments in the domain of Machine Learning (ML) offer great potential to transform AM data into insightful knowledge.

ML techniques offer the ability to discover implicit (formerly unknown) knowledge and identify relationships in large manufacturing data sets, transforming unprecedented volumes of data into actionable and insightful information [10,11]. For AM, ML offers new opportunities to optimize and better control AM processes [12]. In this paper, we explore the state-of-the-art literature on the applications of ML techniques throughout the AM lifecycle.

2. MACHINE LEARNING FOR AM

2.1 Overview of Machine Learning Techniques

Machine learning concerns the construction and study of systems that can automatically learn patterns from data. Models built with ML can be used for prediction, performance optimization, defect detection, classification, regression, or forecasting [10]. The largest factor in determining the effectiveness of ML is the data used to train the ML model. ML models are only as good as the training data has prepared them to be.

ML techniques [13] generally fall into two categories: supervised learning and unsupervised learning. In *supervised learning*, a labeled set of training data provides examples of input values and the corresponding correct output. The ML algorithm trains the model using this labeled dataset, inferring the functional relationship between the input and output domains. Supervised learning can be used for both classification and regression. In *unsupervised learning*, there is no labeled training data set available. Instead, the ML algorithm tries to automatically separate the training dataset into different clusters by grouping parameters in the dataset and identifying target classes. Unsupervised learning is useful for applications such as detecting anomalous conditions. The decision between using a supervised or unsupervised ML approach will depend on perceived benefits for a given scenario.

The typing of supervised and unsupervised models provides a high-level classification in which different ML algorithms can be further categorized. Some popular ML models used for both classification and regression are Support Vector Machines (SVMs) and Neural Networks (NNs). An SVM model identifies hyperplanes that separate the data into different classes. A NN is a computational model that consists of a network of nodes ("neurons") and weighted edges between nodes. NNs are very powerful because they can automatically identify features in the raw data that are needed to make good predictions. These

Razvi, Saadia; Feng, Shaw; Narayanan, Anantha Narayanan; Lee, Yung-Tsun; Witherell, Paul.
"A Review Of Machine Learning Applications In Additive Manufacturing."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE2019), Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

capabilities make NNs very suitable for many AM problems where identifying features in the input data may be difficult.

ML algorithms such as deep learning neural networks are especially useful for very complex tasks such as image and audio processing. Deep learning systems employ several hierarchical layers of processing nodes, which help to identify progressively complex features in the input data. Convolutional neural networks (CNNs) are a special type of deep learning model and are particularly useful for processing image data. A CNN is composed of special processing layers that process image pixels represented as matrices. CNNs progressively extract complex features from an image, such as edges, textures, and shapes, which are used to classify the image, for example as a faulty or good layer in an AM process.

The follow sections introduce how some of the approaches described above have been applied to AM.

2.2. Overview of Literature Survey on ML Applications in AM

In this paper we review research related to ML applications throughout the AM lifecycle. Findings have been gathered from an extensive review of literature published over the last ten years using keyword queries such as "additive manufacturing" and its subcategories [2], coupled with the concepts of "ML." After sifting through hundreds of query results, we analyzed over 50 papers, including journal articles and conference papers. The aims of this review are to 1) identify where ML techniques have been successfully applied in the AM lifecycle, and 2) summarize and organize findings from the existing state-of-the-art research in this domain so that new opportunities can be identified.

Figure 1 categorizes the AM design-to-product transformation cycle based on decision support needs and ML opportunities. Here, we have focused on the following four categories: 1) Design, 2) Process and Performance Optimization, 3) In-Situ Process Monitoring and Control, and 4) Inspection, Testing and Validation. In each of these categories, we focused on a few functions that are currently being analyzed using ML by the research community. For example, in the build phase, research on In-situ Process Monitoring and Control has focused on defect detection, machine-condition monitoring, and real time process control. The following sections delve into the main findings of this survey.

3. AM DESIGN

The AM design process can be decomposed into several stages [5]. Several functions within these stages are currently being implemented using a variety of ML techniques: design recommendations, topology and lattice optimization, tolerancing and manufacturability assessment, and material design and selection. Table 2 presents a summary of the ML techniques used to provide Design Decision Support.

3.1. Design Recommendations

Design-recommendation systems using ML have been developed to assist AM designers. Yao et al. [14] developed a hybrid, machine-learning algorithm to provide design feature recommendations and to assist inexperienced designers in the AM conceptual design phase. Their algorithm combines unsupervised learning (hierarchical clustering) with a trained, supervised classifier (support vector machine (SVM)). Furthermore, they indicate a plan to use ontology-based expert systems to represent more complex AM design knowledge.

Table 2. Overview of ML techniques used to provide Design Decision Support

AM Application	ML Technique	Reference
Design feature recommendation	Hierarchical clustering, SVM	Yao et al., 2017 [14]
Part mass, support material and build-time prediction	NN	Murphy et al., 2019 [15]
Build-time prediction	NN	Munguía et al., 2008 [16] Di Angelo and Di Stefano, 2011 [17]
Cost estimation	Dynamic clustering, LASSO, Elastic net regression	Chan et al., 2018 [18]
Topology optimization	Genetic algorithms, NN	Gaynor et al., 2015 [19]
Geometry compensation to counter thermal shrinkage and deformation	Feed-forward NN with back-propagation	Chowdhury and Anand, 2016 [20]
Shape deviation prediction (tolerancing)	Bayesian Inference	Zhu et al., 2018 [21]
Classification of AM powders	CNN, Random Forest Network (RFN) SVM	Ling et al., 2017 [22] DeCost et. Al, 2017 [23]

Researchers are also employing ML techniques to help novice designers predict design for AM (DfAM) attributes such as expected build time and required support structures. Murphy et al. [15] employed 1) an autoencoder NN that was trained to compress and reconstruct voxelized part designs followed by 2) predictive NNs to predict part mass, support material, and build time. Their existing efforts have achieved limited prediction accuracy; consequently, they plan to implement CNNs in the future to improve accuracy by recognizing and representing local geometries such as lattices.

Munguía et al. [16] used an NN to predict build time for Laser Powder Bed Fusion (L-PBF). NN was used for two reasons. First, it can learn and adapt to different cases. Second, it provides accurate estimates regardless of the different types of machine models. These estimates were calculated using only three parameters: Z-height, part volume, and bounding-box volume. Compared to analytical and parametric time estimators, which have prediction errors rates between 20-25 percent, the NN resulted in error rates between 2-15 percent. Similarly, Di Angelo and Di Stefano [17] also implemented a NN-based build-time estimator. However, they used a parametric approach to capture a more complete set of build-time factors that considered both the dimensional and the geometric features of the object. The authors claim that their custom-designed NN, which used eight build-time driving factors, yielded successful results.

Additionally, researchers are using ML techniques to develop cost-estimation frameworks for AM by leveraging the large amounts of available product and production-related data. For example, Chan et al. [18] predicted the cost for a new print

Razvi, Saadia; Feng, Shaw; Narayanan, Anantha Narayanan; Lee, Yung-Tsun; Witherell, Paul.
"A Review Of Machine Learning Applications In Additive Manufacturing."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE2019), Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

job based on historical data from similar parts. They used the similarities in the 3D geometry and printing processes of parts to extract important features from the part geometry. ML algorithms for dynamic clustering, least absolute shrinkage and selection operator (LASSO), and elastic net regression are applied to feature vectors to predict cost based on historical data.

3.2 Topology Optimization

Topology optimization is a more critical problem in AM than traditional, subtractive processes because of the enormous customizability offered by AM processes. Optimization, in this case, usually means selecting the topology that minimizes the total mass of the structure. Gradient-based optimization algorithms, stochastic algorithms such as genetic algorithms, and NNs have all been explored for topology optimization in AM [19].

Chowdhury and Anand [20] developed a geometry-compensation method to counteract thermal deformation in AM parts caused by temperature gradients during AM fabrication. Their methodology uses a back-propagation NN, trained on surface data from the CAD model, to predict the surface of the fabricated part. The trained network can modify the stereolithography (STL) file whenever the CAD surface data for a new part predicts poor surface quality of the final part. The authors successfully demonstrated a reduced error in manufactured parts' conformity to CAD design by using their NN results.

3.3 Tolerancing and Manufacturability Assessment

Zhu et al. [21] proposed a prescriptive, deviation-modelling method coupled with ML techniques to accurately model shape deviations in AM. Bayesian inference is used to estimate geometric deviation patterns by statistical learning from different shape data, thus supporting more accurate tolerancing for AM parts.

In addition to tolerancing, researchers are using ML to assess the manufacturability of AM-designed parts. Balu et al. [24] proposed a deep-learning-based approach for assessing manufacturability. Deep learning is used to learn different Design for Manufacturing (DFM) rules from labeled voxelized CAD models, without additional shape or process information. AM is mentioned as an applicable technology that could benefit from such a deep-learning-based DFM framework.

3.4 Material Classification and Selection

Machine-learning techniques have been explored to uncover knowledge about the fundamental physical, mechanical, electrical, electronic, chemical, biological, and engineering properties of materials [25]. This knowledge is particularly useful for the classification of AM powders. Ling at el. [22] used deep-learning techniques to classify SEM images of AM powders based on the different powder-size distributions. A CNN was used to transform images and extract features. A random forest network (RFN) classifier was used to sort the transformed images into different size distributions.

DeCost et al. [23] developed a feature, detection-and-description algorithm to create micro-structural-scale, image representations of AM powders. The algorithm applied computer-vision techniques to capture the image of the real object. Scale-invariant feature transformations, together with a vector of locally aggregated descriptors, were then used to encode that image into a digital representation. The authors used this encoding approach, over a NN-based representation, due to its strong rotation and scale invariances. This feature is important because AM powder micrographs do not have any natural orientation. After applying the algorithm, the authors used an SVM to classify the various representations into different material systems, with an accuracy greater than 95 percent.

4. AM PROCESS AND PERFORMANCE OPTIMIZATION

A growing field of study is using data-driven analysis to map the complex relationships among process (P) parameters, final material structure (S), properties (P) and performance (P), also known as PSPP, of the AM part [26]. While finite element modeling (FEM) methods have provided some success in mapping complex PSPP relationships, accurately representing AM processes using high fidelity modeling is difficult. Physics-based models are complex, requiring a deep understanding of material properties and the physical laws governing the AM process. Low fidelity models suffer from lack of information about physical properties, specially due to variabilities from machine to machine and material to material [27].

ML techniques have the potential to successfully discover complex PSPP relationships, overcoming many of the limitations associated with the techniques listed above. The gamut of such techniques generally focuses on understanding either process response or performance response [26], by either using data-driven approach, or combining both physics-based and data-driven approaches. Table 3 presents a summary of literature reviewed in this domain.

Table 3. Overview of ML techniques used for AM Process and Performance Optimization

AM Application	ML Technique	Reference
Build precision (deposition height) prediction	Back propagation (BP) NN, LS-SVM	Lu et al., 2010 [28]
Process parameter optimization (melt pool depth and height)	Genetic algorithm, Self-organizing maps	Fathi and Mozaffari, 2014 [29]
Powder spreading prediction	BP NN	Zhang et al., 2017 [30]
Melt pool width prediction	Gaussian Process Regression	Yang et al., 2018 [31]
Material toughness optimization	Self-Consistent Clustering	Yan et al., 2018 [32]
Porosity prediction	RFN	Kappes et al., 2018 [33]
Wear strength prediction	Genetic programming, NN	Garg and Tai, 2014 [34]
Part density prediction	Kriging, Polynomial regression, NN	Yang et al., 2018 [35]

4.1 Data-Driven Approaches to Characterize Process Response

Lu et al. [28] used a variety of ML techniques to monitor responses in a Directed Energy Deposition (DED) process. Specifically, they map the complex, non-linear relationship

Razvi, Saadia; Feng, Shaw; Narayanan, Anantha Narayanan; Lee, Yung-Tsun; Witherell, Paul.
"A Review Of Machine Learning Applications In Additive Manufacturing."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE2019), Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

between DED process parameters - laser power, scanning speed, and feed rate – and one performance response – building precision as measured by deposition height. The authors adapted a back propagation NN (BP NN) with an adaptive, learning rate, and a momentum coefficient algorithm. The modifications accelerated the training time and improved the results.

Similarly, Fathi and Mozaffari [29] developed a data-driven framework for optimizing process parameters in L-PBF. The authors used a bio-inspired, optimization algorithm, called Mutable Smart Bee algorithm, and a fuzzy inference system to relate process parameters to melt-pool depth and layer height. Derived relationships were combined with a non-dominated, sorting, genetic algorithm to optimize process parameters. Additionally, they proposed using an unsupervised, machine-learning approach - known as self-organizing maps - to further post-optimize the process.

4.2 Physics-Based-Simulation Approaches to Characterize Process Response

In lieu of empirical data, another ML approach is creating surrogate models from physics-based simulation data. For example, Zhang et al. [30] used ML to predict powder-spread parameters as a function of spreading speed and surface roughness of the powder bed. They developed a synergistic, multi-step framework combining 1) a Discrete Element Method (DEM) to simulate a powder spreading process with 2) a BP NN to regress between the highly non-linear results obtained from DEM. The result is a powder-spreading process map that can be used by AM operators to manufacture parts with desired surface roughness.

Yang et al. [31] used the results from an L-PBF, single-track, heat-transfer simulation to predict melt-pool width for different combinations of processing conditions. Their prediction approach combines a Dynamic Variance-Covariance Matrix, the kriging method, Gaussian Process Regression, and genetic algorithms to optimize process parameters. Their approach led to a maximum, relative, error magnitude (MREM) less than 0.03 percent and an average, relative, error magnitude (AREM) less than 0.005 percent for the AM case study.

4.3 Combined Approaches to Characterize Performance Response

The process response, together with the raw-material properties and the final-design structure, are critical factors in predicting the performance response of AM-fabricated parts. Yan et al. [32] proposed combining physics-based models, process models, material models, and data-mining techniques to better understand those factors and their relationship to performance. In this case, the performance response was the mechanical toughness of the built part. The authors combined self-consistent clustering analysis with a reduced-order modeling technique to predict the toughness. They did so by mapping the microstructural descriptors to toughness. However, they discovered that ML techniques like Kriging and NN are better suited for evaluating larger databases. They propose to use this discovery in the future for comprehensive modeling of PSP relationships.

Kappes et al. [33] focused on predicting three performance responses for AM-built parts: fraction porosity, median pore-diameter and median pore-spacing. Their goal was to predict responses by combining information/models about the process (L-PBF), the structure (sample position and orientation), and the material (Inconel 718). The authors used an RFN to make those predictions for two reasons. First, RFN is capable of both classification and regression. Second, RFN is insensitive to irrelevant features. These capabilities were important because, in AM, not all processing conditions are consistently important across different processes and materials.

In another approach, Garg and Tai [34] combined genetic programming and NN using the least squares method. This combined model was used to predict the wear strength of aerospace parts produced using Fused Deposition Molding (FDM). The final structure and raw material for each part were known in advance. The process variables were layer thickness, orientation, raster angle, raster width, and air gap. The values of these variables provided the inputs into both the GP and the NN. The authors showed that their combined approach gave better statistical predictions than using a single ML algorithm.

Similarly, Yang et al. proposed a super-metamodeling framework (SMOF) to predict relative density of AM parts as a function of process parameters such as scanning speed, scanning spacing and laser pulse frequency in an L-PBF process [35]. The SMOF was built by aggregating Kriging, polynomial regression, and NN into a weighted composite to improve overall prediction accuracy while being insensitive to dataset variation. The results positively indicated the superiority of SMOF over individual metamodels, with a final AREM of only 5.47 percent.

5. IN-SITU PROCESS MONITORING AND CONTROL

One of the most focused areas of machine-learning applications in AM is in-situ process monitoring and control. In-situ monitoring technologies are rapidly growing; they now include highs-speed optical cameras, thermocouples, pyrometers, and photo-detectors, among other sensors [6]. However, achieving real-time control for AM is still at a nascent stage – despite the streams of “big,” multi-modal, sensing data capable of being collected. This is due to a few reasons. First, it is still unclear which sensor data is most meaningful for implementing control strategies. Second, the “data-fusion” techniques needed to understand all that sensor data do not exist. Finally, the ML techniques needed to analyze that fused data do exist; but, they have only recently been applied in AM.

Nevertheless, by using in-situ data to characterize the current “state” of a part, combined with a priori knowledge of part and process, we can predict the “state” of the final part [6]. Using ML to improve real-time control of AM fabrication processes has a significant potential benefit – post-process-inspection tasks might be reduced – possibly significantly. By moving some of that post-process inspection upstream, as part of the fabrication process, potential defects in the final parts could be detected earlier. This saves inspection time; but, it also saves materials and processing [36].

Current efforts towards using ML to realize the vision of real-time control for AM processes are primarily focused on

Razvi, Saadia; Feng, Shaw; Narayanan, Anantha Narayanan; Lee, Yung-Tsun; Witherell, Paul.
”A Review Of Machine Learning Applications In Additive Manufacturing.”
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE2019), Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

monitoring the state of either the built part, or the AM machine itself. Some elementary work on process control has also been done. Table 4 provides a summary of literature reviewed in this domain.

Table 4. Overview of ML techniques used for in-situ process monitoring and control

AM Application	ML Technique	Reference
Part Defect Detection and Prediction		
Porosity detection	SVM, k-Nearest Neighbors (k-NN), feed forward NN	Imani et al., 2018 [37]
Quality of fusion and defect detection	Bayesian classifier	Aminzadeh and Kurfess, 2018 [38]
Anomaly detection and classification	Bag-of-keypoints (words), K-means unsupervised clustering, CNN	Scime and Beuth, 2018 [39,40]
Melt pool features and spatter detection	SVM, CNN	Zhang et al., 2018 [41]
Defect detection and classification with acoustic emissions	Spectral CNN	Shevchik et al., 2018 [42]
	Probabilistic graph-based deep belief networks	Ye et al., 2018 [43]
Fault detection from multi-sensor data	Support Vector Data Description (SVDD)	Grasso et al., 2018 [44]
Quality monitoring using heterogeneous sensors in FDM	Bayesian Dirichlet process, Evidence Theory, NN, Naïve Bayes clustering, SVM, Quadratic discriminant analysis	Rao et al., 2015 [45]
Defect detection for L-PBF using in-situ images coupled with ex-situ CT scans	SVM, NN SVM ensemble classifier	Petrich et al., 2017 [46] Gobert et al., 2018 [47]
Machine-Condition Monitoring		
Machine-condition monitoring	k-NN, Bayes Classifier, NN, SVM	Uhlmann et al., 2017 [48]
FDM machine-condition monitoring using acoustic emissions	SVM, K-means clustering, Hidden semi-Markov model	Wu et al., 2015 [49][50][51]
Process Control		
PID process control for FDM	SVM	Liu et al., 2017 [52]
Image-guided process control for L-PBF	Markov Decision Process	Yao et al., 2018 [53]

5.1. Part Defect Detection and Prediction

AM parts can have several different types of defects including porosity, poor surface finish, layer delamination, cracking, and geometric distortion, to name a few [54]. Detecting defects is important to identifying failed builds and predicting the final properties of the part.

5.1.1 Defect Detection with Visual Data

Imani et al. [37,55] presented a qualify-as-you-build model where ML techniques use real-time sensor data to identify process conditions that are likely to cause porosity. The authors analyzed the relationship between laser power, hatch spacing, and velocity, on the size, frequency and location of pores in parts produced through L-PBF. Statistical features are extracted from layer-by-layer in-situ images. These features are subsequently classified by ML techniques like SVM, k-NN, and feed forward NN to identify process conditions most likely to produce pores.

Aminzadeh and Kurfess [38] developed an online monitoring system, using computer vision and Bayesian inference, to inspect both the porosity and the quality of parts in metal L-PBF. They created a labeled dataset of defective and non-defective features from in-situ camera images of each layer. They extracted frequency-domain features from those images and used a Bayesian classifier to identify of defective vs non-defective parts.

Instead of using layer wise images of the powder after laser interaction, Scime and Beuth [39] used computer vision and ML techniques to detect and classify anomalies and flaws in the powder prior to fusion. They investigated six different types of powder bed anomalies captured in labeled images from an L-BPF machine. The bag-of-keypoints ML technique used to detect and classify anomalies was able to detect the presence of an anomaly in 89 percent of cases, with 95 percent accuracy in correctly identifying the type of anomaly. Separately, the authors showed that accuracy can be further improved by implementing a multi-scale CNN for autonomous anomaly detection and classification [40].

Zhang et al. [41] used Principal Component Analysis (PCA) with SVM to enable using CNN to recognize features in the laser melting process. Features include melt pool, spatter, plume, and anomalies. The accuracy is reported to be 92.7 percent.

5.1.2 Defect Detection with Acoustic Data

Acoustic emissions (AE) have also been used for defect detection. AE sensors are non-intrusive to the build process and provide high throughput for real-time monitoring. Ye et al. [43] developed a method of analyzing acoustic signals with deep belief networks (DBN) to detect defects in the L-PBF process. Temperature changes from melting to solidification create variations in the acoustic signals. The authors trained a DBN to recognize defects based on the categorizations of balling, keyholing, and cracking, using the sparking sound spectrum in the time domain and the signal power spectral density in the frequency domain.

Shevchik et al. [42] investigated the use of AE combined with CNN to detect various defects due to lack of fusion. The authors used a fiber Bragg grating acoustic sensor to detect the airborne AE signals, generated from the melting, sparking, spattering, and solidification processes. The signals collected in the time domain are transformed to the frequency domain using the wavelet packet transform, an extension of the traditional wavelet transform. The Spectral CNN, an extension of CNN with improved efficiency in classification and regression, is used to recognize features in the frequency domain that correspond to defects in the L-PBF process. The confidence level in SCNN is between 83 to 89 percent, according to the authors.

5.1.3 Defect Detection with Multi-Sensor Data

As aforementioned, data gathered from in-situ monitoring of AM processes is highly varied. Registering and fusing together data from multiple sensors provides a rich context for

Razvi, Saadia; Feng, Shaw; Narayanan, Anantha Narayanan; Lee, Yung-Tsun; Witherell, Paul.
"A Review Of Machine Learning Applications In Additive Manufacturing."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE2019), Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

fault detection. Therefore, a growing area of research involves multi-sensor data fusion for process monitoring and control.

Grasso et al. [44] explored data fusion methodologies to combine in-situ data from multiple sensors embedded in Electron Beam PBF systems. The Support Vector Data Description (SVDD) ML technique is used to classify in-control vs. out-of-control process signals. The SVDD automatically detects faults and process errors that can be related to the stability of embedded signals from multiple sensor data streams. The limitation of their approach is that it applies only to serial production of the same product.

Rao et al. [45] fused data from a heterogeneous sensor suite as part of an online-monitoring system for FDM. The suite comprises of thermocouples, accelerometers, an infrared temperature sensor, and a real-time, miniature, video borescope. Process failures (such as nozzle clog) are detected from the fused sensor data using the non-parametric Bayesian Dirichlet process mixture model and evidence theory, achieving a prediction accuracy of up to 85 percent. In comparison, existing approaches, such as probabilistic NN, Naïve Bayes clustering, and SVM had poorer performance.

Petrich et al. [46] and Gobert et al. [47] used multi-sensor data fusion to detect discontinuity defects – such as pores, over-heating areas, and unmolten powders – in L-PBF. They merged together homogenous sensor data (eight sets of layer-wise images of the powder bed under varying lighting conditions, pre- and post-sintering) with heterogenous sensor data (post build CT scans). Ground-truth labels (anomalous or normal) extracted from the CT scans were used to train NN and SVM [46], as well as SVM ensemble classifiers [47] to detect defects directly from images. Ensemble classifiers can analyze multiple images under different lighting conditions with a high classification accuracy (85 percent) as compared to classification using images from only a single lighting condition (65 percent accuracy).

5.2. Machine-Condition Monitoring

Another approach to in-situ monitoring is observing the machine logs or build condition instead of monitoring the part. Clustering techniques can classify features extracted from machine logs and identify normal or problematic build states [48]. In a series of papers, Wu et al. [49–51] developed an approach for FDM machine condition monitoring using AE data to identify normal and abnormal machine states. They extracted time- and frequency-domain features from the data and used a variety of ML algorithms (SVM with radial bias function kernel [49], K-means clustering [50], and hidden semi-Markov model [51]) to classify normal vs. abnormal machine-condition states. Their monitoring method can be used as a diagnostic tool to identify failure states such as material runout or filament breakage.

5.3. Process Control

Liu et al. [52] developed an online closed-loop controller for FDM. Their control architecture consists of 1) real-time image acquisition, 2) a tool for image analysis, and 3) a Proportional-Integral-Derivative (PID) controller for closed-loop control. They identified two types of defects, overfill and underfill, at different severities. After extracting textural features from the image data collected from a microscope, they used SVM to differentiate those features into two groups: normal and defective. Then they used another SVM to identify the severity of defects. The PID controller used the results of that analysis to modify the feed rate to mitigate each type of defect.

Yao et al. [53] developed a smart, closed-loop optimal control system for L-PBF. They used multifractal analysis to estimate the defect condition of each layer, and then predicted the future evolution of defects in following layers. Finally, they modeled the stochastic dynamics of layer-to-layer defect conditions as a Markov decision process for deriving an optimal control policy.

6. INSPECTION, TESTING AND VALIDATION

ML techniques are used for final AM part inspection and validation. The focus is primarily on surface metrology, and defect detection and classification using ex-situ measurements, such as X-CT data. Table 5 presents an overview of the literature reviewed in this domain (excluding X-CT).

Table 5. Overview of ML techniques used for post-process inspection and validation

AM Application	ML Technique	Reference
Classification of dimensional variation from laser scanned 3D point cloud data	Sparse representation, k-NN, NN, Naïve Bayes SVM, Decision tree	Tootani et al., 2017 [56]
Defect detection (porosity)	Augmented layer-wise spatial log Gaussian Cox process (ALS-LGCP)	Liu et al., 2018 [57]

Tootooni et al. [56] developed a new method to classify dimensional variations in parts made with FDM based on spectral graph theory. They used Laplacian Eigenvalues as extracted features from laser-scanned 3D point cloud data, followed by supervised ML techniques to classify dimensional variation, including sparse representation, k-NN, NN, naïve Bayes, SVM, and decision tree. The sparse representation technique provided the highest classification accuracy (F-score > 95 percent). Their approach requires a priori knowledge of the part for training, thus limiting applications to other parts.

Liu at al. [57] proposed an augmented layer-wise spatiotemporal log Gaussian Cox process (ALS-LGCP) model to quantify the spatial distribution of pores within each layer of an AM part and track sequential evolution across layers. They applied the ALS-LGCP to binder-jetted parts, and used Bayesian predictive analytics to predict porosity prone areas in successive layers, achieving statistical fidelity approaching 85 percent.

Senin and Leach [58] developed a smart information-rich surface metrology technique using multi-sensor data fusion and ML. They identified AM as an example where advanced measurement techniques are needed due to complex geometries and lack of uniform material properties.

7. CONCLUSION

This paper presents a detailed review of ML applications throughout the AM design-to-product transformation cycle. We have categorized the literature based on the applications as they

Razvi, Saadia; Feng, Shaw; Narayanan, Anantha Narayanan; Lee, Yung-Tsun; Witherell, Paul.
"A Review Of Machine Learning Applications In Additive Manufacturing."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE2019), Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

pertain to the different phases in the AM product lifecycle. With most of the reviewed research published in 2017 or later, the ML methods identified throughout this paper are the beginnings of what is sure to be a growing effort of ML applications for AM. We observed why ML methods are well suited to solve problems in the AM domain and which methods are most commonly being used. To date, ML for AM research has been opportunistic, where researchers have identified areas rich with data, such as in in-situ process monitoring and control. The high dimensionality and complexity of AM data makes it well-suited for popular ML algorithms. For instance, supervised learning techniques, such as NN and SVM, are most popular due to the availability of labeled datasets. This paper lays a foundation for a more methodical approach to ML for AM moving forward.

While ML techniques are rapidly being adopted into AM applications, there are many opportunities for improved future applications. For instance, unsupervised learning techniques are not as widely adopted. However, with the increasing amounts of unlabeled AM datasets, these techniques are likely to become more popular and thus should be further investigated. Alternatively, as ML algorithms require training data, increased interest in ML for AM will lead to new approaches for supervised ML.

ML models are very poor at diagnosing conditions that have not been previously encountered. This limitation puts an emphasis on collecting data for training by creating scenarios that will address a wide range of operating conditions and dimensionality space. A major challenge in the maturation of ML for AM is the lack of availability of accurate, accessible, and extensive databases for AM processes, products, and materials [26]. While each build can generate terabytes of data, there is a lack of standard practices for handing datasets characterized by high volume and velocity in real time.

Absence of a common data structure, and standard methods for data integration and fusion, prevents rich, multifaceted, data-driven analysis. Furthermore, generating exemplar data via experimentation is difficult and expensive. Even if data is available, poor quality of data makes it unsuitable for ML algorithms. Low resolution of in-situ optical data, limited fields of view, and high temporal load result in poor quality data sets [6]. This hinders feature selection for ML algorithms. The development of feature libraries for AM feature characterization would help address some of the current challenges that make it difficult to select a suitable ML algorithm compatible with the available data.

8. FUTURE WORK

As reviewed in this paper, there already are many AM applications that are benefitting from ML techniques. However, even more applications areas remain unexplored. For instance, in the domain of AM design, deep learning techniques could be used to train on voxelized CAD models to make better predictions of DfAM attributes such as part mass, support structures, and build time. The in-situ, monitoring-and-control domain could benefit from the advantages of deep learning techniques for use in fault detection and build failures. CNN, for example, can detect and classify both macroscopic and microscopic faults using layer-wise, optical-sensor data. Moving forward, potential opportunities like these will continue to be identified.

Identifying new opportunities in the AM lifecycle is simply a precursor to the data challenges that will arise when seeking to take advantage of these opportunities. For instance, further research is needed for in-situ data sensor fusion. The fusion of thermal, acoustic, optical and other build environmental data can create a more holistic, reliable and accurate information source for real-time defect detection and correction with feedback control. Other opportunities include using ML to build models correlating in-situ and ex-situ data, such as IR videos with NDE X-CT data. Such an approach could enable the “qualify-as-you-build” goal for AM and reduce dependence on post build NDE qualification. As new AM data sets continue to emerge so will new opportunities to leverage ML techniques to improve the fabrication of AM parts.

ACKNOWLEDGEMENTS AND DISCLAIMER

Certain commercial systems are identified in this paper. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology (NIST); nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST or any other supporting U.S. government or corporate organizations.

REFERENCES

[1] Berman, B., 2012, “3-D Printing: The New Industrial Revolution,” Bus. Horiz., **55**(2), pp. 155–162.

[2] ASTM International, 2015, “Standard Terminology for Additive Manufacturing – General Principles – Terminology,” ASTM ISO/ASTM52900-15.

[3] Gao, W., Zhang, Y., Ramanujan, D., Ramani, K., Chen, Y., Williams, C. B., Wang, C. C. L., Shin, Y. C., Zhang, S., and Zavattieri, P. D., 2015, “The Status, Challenges, and Future of Additive Manufacturing in Engineering,” Comput.-Aided Des., **69**, pp. 65–89.

[4] *Wohlers Associates, Inc., 2016, Wohlers Report 2016: Additive Manufacturing and 3D Printing State of the Industry, Wohlers Associates, Fort Collins, CO.*

[5] Kim, D. B., Witherell, P., Lipman, R., and Feng, S. C., 2015, “Streamlining the Additive Manufacturing Digital Spectrum: A Systems Approach,” Addit. Manuf., **5**, pp. 20–30.

[6] Everton, S. K., Hirsch, M., Stravroulakis, P., Leach, R. K., and Clare, A. T., 2016, “Review of In-Situ Process Monitoring and in-Situ Metrology for Metal Additive Manufacturing,” Mater. Des., **95**, pp. 431–445.

[7] Mani, M., Lane, B., Donmez, A., Feng, S., Moylan, S., and Fesperman, R., 2015, "Measurement Science Needs for Real-Time Control of Additive Manufacturing Powder Bed Fusion Processes", NIST IR 8036, National Institute of Standards and Technology, Gaithersburg, MD, USA.

Razvi, Saadia; Feng, Shaw; Narayanan, Anantha Narayanan; Lee, Yung-Tsun; Witherell, Paul.
”A Review Of Machine Learning Applications In Additive Manufacturing.”
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE2019), Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

[8] Witherell, P., 2018, “Emerging Datasets and Analytics Opportunities in Metals Additive Manufacturing,” *Direct Digital Manufacturing Conference*, Berlin, Germany.
[9] Dehoff, R., 2015, “In-Situ Process Monitoring and Big Data Analysis for Additive Manufacturing of Ti-6All-4V,” *Titanium USA Conference 2015*, ITA, Orlando, FL, USA
[10] Wuest, T., Weimer, D., Irgens, C., and Thoben, K.-D., 2016, “Machine Learning in Manufacturing: Advantages, Challenges, and Applications,” Prod. Manuf. Res., **4**(1), pp. 23–45.
[11] Wang, J., Ma, Y., Zhang, L., Gao, R. X., and Wu, D., 2018, “Deep Learning for Smart Manufacturing: Methods and Applications,” J. Manuf. Syst., **48**, pp. 144–156.
[12] Yang, J., Chen, Y., Huang, W., and Li, Y., 2017, “Survey on Artificial Intelligence for Additive Manufacturing,” *2017 23rd International Conference on Automation and Computing (ICAC)*, IEEE, Huddersfield, United Kingdom, pp. 1–6.
[13] Hastie, T., Tibshirani, R., and Friedman, J., 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.
[14] Yao, X., Moon, S. K., and Bi, G., 2017, “A Hybrid Machine Learning Approach for Additive Manufacturing Design Feature Recommendation,” Rapid Prototyp. J., **23**(6), pp. 983–997.
[15] Murphy, C., Meisel, N., Simpson, T. W., and McComb, C., 2018, “Predicting Part Mass, Required Support Material, and Build Time via Autoencoded Voxel Patterns” [Online]. Available: engrxiv.org/8kne7
[16] Munguía, J., Ciurana, J., and Riba, C ., 2008, "Neural-Network-Based Model for Build-Time Estimation in Selective Laser Sintering", *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, **223**(8), 995–1003.
[17] Di Angelo, L., and Di Stefano, P., 2011, “A Neural Network-Based Build Time Estimator for Layer Manufactured Objects,” Int. J. Adv. Manuf. Technol., **57**(1–4), pp. 215–224.
[18] Chan, S. L., Lu, Y., and Wang, Y., 2018, “Data-Driven Cost Estimation for Additive Manufacturing in Cybermanufacturing,” J. Manuf. Syst., **46**, pp. 115–126.
[19] Gaynor, A. T., 2015, “Topology Optimization Algorithms for Additive Manufacturing,” Ph.D. thesis, Department of Civil Engineering, Johns Hopkins University.
[20] Chowdhury, S., and Anand, S., 2016, “Artificial Neural Network Based Geometric Compensation for Thermal Deformation in Additive Manufacturing Processes,” *ASME 2016 International Manufacturing Science and Engineering Conference MSEC2016*, ASME, Blacksburg, Virginia, USA.
[21] Zhu, Z., Anwer, N., Huang, Q., and Mathieu, L., 2018, “Machine Learning in Tolerancing for Additive Manufacturing,” CIRP Ann.
[22] Ling, J., Hutchinson, M., Antono, E., DeCost, B., Holm, E. A., and Meredig, B., 2017, “Building Data-Driven Models with Microstructural Images: Generalization and Interpretability,” [Online]. Available: ArXiv171100404 Cond-Mat.
[23] DeCost, B. L., Jain, H., Rollett, A. D., and Holm, E. A., 2017, “Computer Vision and Machine Learning for Autonomous Characterization of AM Powder Feedstocks,” JOM, **69**(3), pp. 456–465.
[24] Balu, A., Lore, K., Gavin, Y., Krishnamurthy, A., and Sarkar, S., 2016, “A Deep 3D Convolutional Neural Network Based Design for Manufacturability Framework,” [Online]. Available: ArXiv:1612.02141v1 [cs.CV].
[25] Ramakrishna, S., Zhang, T.-Y., Lu, W.-C., Qian, Q., Low, J. S. C., Yune, J. H. R., Tan, D. Z. L., Bressan, S., Sanvito, S., and Kalidindi, S. R., 2018, “Materials Informatics,” J. Intell. Manuf.
[26] Smith, J., Xiong, W., Yan, W., Lin, S., Cheng, P., Kafka, O. L., Wagner, G. J., Cao, J., and Liu, W. K., 2016, “Linking Process, Structure, Property, and Performance for Metal-Based Additive Manufacturing: Computational Approaches with Experimental Support,” Comput. Mech., **57**(4), pp. 583–610.
[27] Baturynska, I., Semeniuta, O., and Martinsen, K., 2018, “Optimization of Process Parameters for Powder Bed Fusion Additive Manufacturing by Combination of Machine Learning and Finite Element Method: A Conceptual Framework,” Procedia CIRP, **67**, pp. 227–232.
[28] Lu, Z. L., Li, D. C., Lu, B. H., Zhang, A. F., Zhu, G. X., and Pi, G., 2010, “The Prediction of the Building Precision in the Laser Engineered Net Shaping Process Using Advanced Networks,” Opt. Lasers Eng., **48**(5), pp. 519–525.
[29] Fathi, A., and Mozaffari, A., 2014, “Vector Optimization of Laser Solid Freeform Fabrication System Using a Hierarchical Mutable Smart Bee-Fuzzy Inference System and Hybrid NSGA-II/Self-Organizing Map,” J. Intell. Manuf., **25**(4), pp. 775–795.
[30] Zhang, W., Mehta, A., Desai, P. S., and III, C. F. H., “Machine Learning Enabled Powder Spreading Process Map for Metal Additive Manufacturing (AM),” p. 15.
[31] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., Witherell, P. W., and Lopez, F., 2018, “Dynamic Metamodeling for Predictive Analytics in Advanced Manufacturing,” Smart Sustain. Manuf. Syst., **2**(1).
[32] Yan, W., Lin, S., Kafka, O. L., Lian, Y., Yu, C., Liu, Z., Yan, J., Wolff, S., Wu, H., Ndip-Agbor, E., Mozaffar, M., Ehmann, K., Cao, J., Wagner, G. J., and Liu, W. K., 2018, “Data-Driven Multi-Scale Multi-Physics Models to Derive Process–Structure–Property Relationships for Additive Manufacturing,” Comput. Mech., **61**(5), pp. 521–541.
[33] Kappes, B., Moorthy, S., Drake, D., Geerlings, H., and Stebner, A., 2018, “Machine Learning to Optimize Additive Manufacturing Parameters for Laser Powder Bed Fusion of Inconel 718,” *9th International Symposium on Superalloy 718 & Derivatives: Energy,*

Razvi, Saadia; Feng, Shaw; Narayanan, Anantha Narayanan; Lee, Yung-Tsun; Witherell, Paul.
"A Review Of Machine Learning Applications In Additive Manufacturing."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE2019), Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

Aerospace, and Industrial Applications, Springer International Publishing, pp. 595–610.
[34] Garg, A., and Tai, K., 2014, "An Ensemble Approach of Machine Learning in Evaluation of Mechanical Property of the Rapid Prototyping Fabricated Prototype," Appl. Mech. Mater., **575**, pp. 493–496.
[35] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., and Lu, Y., 2018, "A Super-Metamodeling Framework to Optimize System Predictability," *ASME 2018 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, ASME, Quebec City, Canada.
[36] Sarah Anderson Goehrke, 2017, "Metal 3D Printing with Machine Learning: GE Tells Us About Smarter Additive Manufacturing," 3DPrintcom Voice 3D Print. Addit. Manuf.
[37] Imani, F., Montazeri, M., Gaikwad, A., Rao, P., Yang, H., and Reutzel, E., 2018, "Layerwise In-Process Quality Monitoring in Laser Powder Bed Fusion," *ASME 2018 13th International Manufacturing Science and Engineering Conference*, ASME, College Station, USA.
[38] Aminzadeh, M., and Kurfess, T. R., 2018, "Online Quality Inspection Using Bayesian Classification in Powder-Bed Additive Manufacturing from High-Resolution Visual Camera Images," J. Intell. Manuf.
[39] Scime, L., and Beuth, J., 2018, "Anomaly Detection and Classification in a Laser Powder Bed Additive Manufacturing Process Using a Trained Computer Vision Algorithm," Addit. Manuf., **19**, pp. 114–126.
[40] Scime, L., and Beuth, J., 2018, "A Multi-Scale Convolutional Neural Network for Autonomous Anomaly Detection and Classification in a Laser Powder Bed Fusion Additive Manufacturing Process," Addit. Manuf., **24**, pp. 273–286.
[41] Zhang, Y., Hong, G. S., Ye, D., Zhu, K., and Fuh, J. Y. H., 2018, "Extraction and Evaluation of Melt Pool, Plume and Spatter Information for Powder-Bed Fusion AM Process Monitoring," Mater. Des., **156**, pp. 458–469.
[42] Shevchik, S. A., Kenel, C., Leinenbach, C., and Wasmer, K., 2018, "Acoustic Emission for in Situ Quality Monitoring in Additive Manufacturing Using Spectral Convolutional Neural Networks," Addit. Manuf., **21**, pp. 598–604.
[43] Ye, D., Hong, G. S., Zhang, Y., Zhu, K., and Fuh, J. Y. H., 2018, "Defect Detection in Selective Laser Melting Technology by Acoustic Signals with Deep Belief Networks," Int. J. Adv. Manuf. Technol., **96**(5–8), pp. 2791–2801.
[44] Grasso, M., Gallina, F., and Colosimo, B. M., 2018, "Data Fusion Methods for Statistical Process Monitoring and Quality Characterization in Metal Additive Manufacturing," Procedia CIRP.
[45] Rao, P. K., Liu, J. (Peter), Roberson, D., Kong, Z. (James), and Williams, C., 2015, "Online Real-Time Quality Monitoring in Additive Manufacturing Processes Using Heterogeneous Sensors," J. Manuf. Sci. Eng., **137**(6).
[46] Petrich, J., Gobert, C., Phoha, S., Nassar, A. R., and Reutzel, E. W., 2017, "Machine Learning for Defect Detection for PBFAM Using High Resolution Layerwise Imaging Coupled with Post-Build CT Scans," *28th Annual International Solid Freeform Fabrication Symposium,* Austin, TX, USA.
[47] Gobert, C., Reutzel, E. W., Petrich, J., Nassar, A. R., and Phoha, S., 2018, "Application of Supervised Machine Learning for Defect Detection during Metallic Powder Bed Fusion Additive Manufacturing Using High Resolution Imaging.," Addit. Manuf., **21**, pp. 517–528.
[48] Uhlmann, E., Pontes, R. P., Laghmouchi, A., and Bergmann, A., 2017, "Intelligent Pattern Recognition of a SLM Machine Process and Sensor Data," Procedia CIRP, **62**, pp. 464–469.
[49] Wu, H., Wang, Y., and Yu, Z., 2015, "In Situ Monitoring of FDM Machine Condition via Acoustic Emission," Int. J. Adv. Manuf. Technol.
[50] Wu, H., Yu, Z., and Wang, Y., 2016, "A New Approach for Online Monitoring of Additive Manufacturing Based on Acoustic Emission," *Volume 3: Joint MSEC-NAMRC Symposia*, ASME, Blacksburg, Virginia, USA.
[51] Wu, H., Yu, Z., and Wang, Y., 2017, "Real-Time FDM Machine Condition Monitoring and Diagnosis Based on Acoustic Emission and Hidden Semi-Markov Model," Int. J. Adv. Manuf. Technol., **90**(5–8), pp. 2027–2036.
[52] Liu, C., Roberson, D., and Kong, Z., 2017, "Textural Analysis-Based Online Closed-Loop Quality Control for Additive Manufacturing Processes," *2017 Industrial and Systems Engineering Conference*, Blacksburg, VA, USA.
[53] Yao, B., Imani, F., and Yang, H., 2018, "Markov Decision Process for Image-Guided Additive Manufacturing," IEEE Robot. Autom. Lett., **3**(4), pp. 2792–2798.
[54] Sames, W. J., List, F. A., Pannala, S., Dehoff, R. R., and Babu, S. S., 2016, "The Metallurgy and Processing Science of Metal Additive Manufacturing," Int. Mater. Rev., **61**(5), pp. 315–360.
[55] Imani, F., Yao, B., Chen, R., Rao, P., and Yang, H., 2018, "Fractal Pattern Recognition of Image Profiles for Manufacturing Process Monitoring and Control," P*roceedings of the ASME 2018 13th International Manufacturing Science and Engineering Conference*, ASME, College Station, TX, USA.
[56] Samie Tootooni, M., Dsouza, A., Donovan, R., Rao, P. K., Kong, Z. (James), and Borgesen, P., 2017, "Classifying the Dimensional Variation in Additive Manufactured Parts From Laser-Scanned Three-Dimensional Point Cloud Data Using Machine Learning Approaches," J. Manuf. Sci. Eng., **139**(9).
[57] Liu, J. (Peter), Liu, C., Bai, Y., Rao, P., Williams, C., and Kong, Z. (James), 2018, "Layer-Wise Spatial Modeling of Porosity in Additive Manufacturing," IISE Trans.
[58] Senin, N., and Leach, R., 2018, "Information-Rich Surface Metrology," Procedia CIRP.

Razvi, Saadia; Feng, Shaw; Narayanan, Anantha Narayanan; Lee, Yung-Tsun; Witherell, Paul.
"A Review Of Machine Learning Applications In Additive Manufacturing."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE2019), Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

Proceedings of the ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2019
August 18-21, 2019, Anaheim, USA

IDETC2019-98429

USING SEMANTIC FLUENCY MODELS IMPROVES NETWORK RECONSTRUCTION ACCURACY OF TACIT ENGINEERING KNOWLEDGE

Thurston Sexton*
Systems Integration Division
Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, Maryland 20871
Email: thurston.sexton@nist.gov

Mark Fuge
Dept. of Mechanical Engineering
University of Maryland
College Park, Maryland 20742
Email: fuge@umd.edu

ABSTRACT

Human- or expert-generated records that describe the behavior of engineered systems over a period of time can be useful for statistical learning techniques like pattern detection or output prediction. However, such data often assumes familiarity of a reader with the relationships between entities within the system—that is, knowledge of the system's structure. This required, but unrecorded "tacit" knowledge makes it difficult to reliably learn patterns of system behavior using statistical modeling techniques on these written records. Part of this difficulty stems from a lack of good models for how engineers generate written records of a system, given their expertise, since they often create such records under time pressure using shorthand notation or internal jargon. In this paper, we model the process of maintenance work order creation as a modified semantic fluency task, to build a probabilistic generative model that can uncover underlying relationships between entities referenced within a complex system. Compared to more traditional similarity-metric-based methods for structure recovery, we directly model a possible cognitive process by which technicians may record work-orders. Mathematically, we represent this as a censored local random walk over a latent network structure representing tacit engineering knowledge. This allows us to recover implied engineering knowledge about system structure by processing written records. Additionally, we show that our model leads to improved generative capabilities for synthesizing plausible data.

1 INTRODUCTION

Due in part to an explosion of interest in statistical modeling techniques, specifically machine learning (ML), much recent effort has been devoted to using various forms of engineering data for training these models. These models, trained on historical engineering data to detect patterns of classification, fault detection, performance estimates, etc., promise to reliably automate many of these labor-intensive tasks, freeing the time of designers and maintainers for more high-level decisions. However, in technical fields like engineering, the available historical data is often difficult to use directly — the experts creating it in the past generally assumed it would be read and adapted by colleagues or experts in their own field. This causes analysts to, quite often, lack the information needed to appropriately represent and process this data. One cannot simply use, *e.g.*, written lab notebooks, technical reports, or maintenance work-orders (MWOs) as is, taking them at face value: words and concepts with more general meaning to the layman will have domain-specific technical application that must be accounted for if a statistical model is to learn a robust representation of the semantic space. In this paper, our goal is to infer how the original data creators/experts structure their own knowledge about the problem at hand. This "structured knowledge" can then be used create more reliable models for engineering learning tasks.

This paper presents initial techniques to automatically infer key parts of this tacit structured knowledge, and explores a mechanism to extract it from observations/historical records written

*Address all correspondence to this author.

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

by human experts. To do this, we frame the act of recording engineering events as a type of memory recall, which we assume occurs within a broader "network" of system relationships that structure the expert's knowledge about a system's behaviors (but that we do not have direct access to and thus must infer through examples). Specifically, we show that:

1. By explicitly modeling work-order generation as non-Markovian memory recall over learned object relationships, we can more accurately recover those relationships than by using more traditional token similarity measures, and subsequently,
2. learning such relationships provides a generative model of each object's conditional relationships in the form of a graph, for which performing a random walk from points of interest (*e.g.*, a Failed part) will synthesize more realistic new data.

We demonstrate this on two examples of maintenance work orders: (1) synthetically generated work orders from real-world engineering systems with a known ground-truth structures; and (2) actual maintenance work orders from an excavator. In both cases, we show that by building a probabilistic model that accounts for (and subsequently learns) how experts structure their implicit knowledge of a domain, one can often achieve significantly better performance (as measured by standard information retrieval metrics) than existing methods of structure recovery.

2 RELATED WORK

Using data to infer the underlying structure of a complex system is a long-standing goal within both systems engineering and other domains that depend upon accurate network recovery, such as: biological systems and disease transmission vector modeling [1, 2]; uncovering economic interactions and social networks [3, 4]; inferring physical models by learning governing equations [5, 6]; or even description generation in computer vision, and quantifying how humans reason about belonging and causality in ambiguous images or contexts [7, 8]. For written (text-based) documents specifically, we can group major methods to perform structure recovery from unstructured written documents into roughly three camps: (1) prescriptive rule definition, (2) training statistical models (NLP), and (3) "folksonomies" and tag-based crowdsourcing.

2.1 Prescriptive Rules

The most straight-forward way to make tacit knowledge computable is to explicitly design the relationships as they are assumed to exist. An expert (or set of experts) define what objects are allowed to exist in the domain, and how those concepts relate to each other. These rules are then mapped onto the observed data, similar to constructing a thesaurus. This manually constructed rule-set can take the form of ontologies, *e.g.*, but they are always structured representations formed from from mixtures of domain expertise and example data, which can then be used to parse remaining data, and restrict the format of future data. For example, ISO-15926 defines a data model [9, 10] with which one can constrain engineering records to have precise, unambigious meanings, and later work built on the standard construct ontologies with which to reason over these meanings and their relationships [9, 11–13].

In practice, however, a particular domain or data-set will not have existing, applicable ontologies or data structures, and time investment needed to create them for sufficiently generalized usage is commonly out-of-scope for analysts to dedicate. Some work has been done to automate this process [14], but such techniques generally require us to rely on language-specific syntactical rules (*i.e.*, grammar). Data-entry errors and shorthand are ubiquitous in technical records, where grammar is often low-priority if system-familiarity is assumed. In these cases, sophisticated systems of rules are still often developed, potentially with reduced formalism or scope, taking the form of keyword recognition and filtering rules to find a priori "useful" patterns for analysis [15, 16] In engineering design, similar manually-created rule-sets that define concept relationships are involved in constructing Design Structure Matrices (DSMs), which are often derived from expert input or technical/project documents [17–19]. Regardless, this paper assumes that the need for low-cost, low effort estimates of a system's "rules" is not met by requiring a designer to manually intervene.

2.2 Natural Language Processing

Rather than build patterns manually, natural language processing (NLP) often deals with the use of significant quantities of text to discover latent patterns automatically. This requires finding mathematical representations of text, like "bag-of-words" weightings [20], topic models [21, 22], or semantic vector embeddings [23, 24]. These transformations enable the use of text-based documents in statistical models that can, for instance, train a classifier to automate labeling of work orders [25]. The success of this approach is fundamentally linked to the notion that supervised ML models use *labeled training data* to learn these patterns, and the quality of the model increases with the amount of available labeled data-points — using this approach with few labels presents a problem of diminishing returns. Time saved by automating document classification scales with the amount of time spent labeling document classifications.

This trade-off is problematic in highly technical and jargon-filled domains, where existing models from more generalized training sets cannot be easily or reliably transitioned. In addition, the actual patterns being "learned" are quite often difficult to interpret and use for humans [26], stemming from the so-called "black box" nature of these models, despite an inherent need to

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

justify our engineering decisions with evidence-based reasoning. This paper puts forward an unsupervised model, able to function with few training samples, but only as a stepping stone in the process toward encoding the types of prescriptive knowledge that can be used to communicate and train future operators/designers.

2.3 Folksonomies

In contexts where dedicated annotation labor can be difficult to secure, significant research has been done to present less restrictions to casual annotators, and understand how natural classification and labeling schemes arise in social communities, *e.g.* online tagging efforts [27]. Tags, a form of multi-label classification, allow concepts to be derived freely in the course of work, where repeated and cross-contexutal usage leads to a naturally-arising set of useful, domain-specific concepts; this is commonly referred to as a *folksonomy*, a portmanteau of "folk" and "taxonomy" [28]. Because folksonomies generally ask users to determine minimal representative labels rather than strict classifications (*i.e.*, tags), each label can be seen in multiple contexts. The predominant way to analyze these tags, then, is by their co-occurrences with each other: intuitively, highly co-occurring tags are considered "similar." [29, 30]. A basic, but commonly used measure of this co-occurrence is the *cosine similarity*: if, over a set of C documents, tag t_k has binary vector $u_k = \{\mathbf{1}_c(t_k) : c \in C\}$, then the cosine similarity s between the binary occurrence vectors of the tags t_1, t_2 is defined as:

$$s(t_1, t_2) = \frac{u_1 \cdot u_2}{\|u_1\| \, \|u_2\|} \tag{1}$$

This measure has seen consistent usage in folskonometric methods to structuring relationships between tagged concepts in useful ways [31–33]. Because various annotators will perceive the importance and relevance of each tag differently in each context, these ambiguities are typically overcome through crowdsourcing, by having large numbers of users tag. This allows a statistical "smoothing" over differences in expertise. However, in the case of technical tags from a few experts, this benefit from large numbers of annotators is not something that we can count on. Additionally, the types of relationship information we might want is not purely statistical/distributional similarity, as experts creating documents will have several core views about what "being related" in their system entails. Consequently, we believe it is important to exploit potential cognitive processes by which these tags might be produced, to enforce a greater degree of information *precision* than typical similarity measures might allow for in their desire for increased information *recall*.

3 MODELING WORK ORDER CREATION

As discussed above, common techniques for discovering structure in human-annotated or natural-language data primarily rely on frequency and co-occurrence information of discrete objects/concepts. These are powerful and easy-to-apply models of speech or the written word, but can miss key causal links implied in the original text, which are difficult to extract this way without significant amounts of data or relevant pre-training. Instead, this paper proposes that by explicitly modeling the conditional dynamics of how humans recall concepts within this data—which, for the purposes of this work will be limited to MWO's—we can extract the conditional relationships between the mentioned objects or concepts that best match what was recorded by the experts.

This section first describes the concept of *semantic fluency*—a existing psychological theory of concept recall—and how that theory relates to the construction of written engineering documents, specifically MWOs in this paper. We then describe a computational method to implement the concept of semantic fluency using Initial-Visit Emitting Random Walks (INVITE) [34]—a probabilistic model of graph walks that is non-Markovian.

3.1 Semantic Fluency

When a technician begins to record a MWO, they try to search their memory for words that represent concepts relevant to the MWO itself. These consist of items, problems that were encountered with some items, and how other items were used to solve these problems [35]. The exact psychological mechanisms by which a person searches through their memory is still an active area of research and has been modeled in various ways. Some recent studies [36] propose that concepts are recalled sequentially by foraging in "semantic patches"—in brief, that humans sequentially recall concepts that are "near" each other in some person-specific semantic space built through experience.

Specifically, these patches are thought of as existing in a high-dimensional concept-space,[1] and the likelihood that some concept is recalled next is based on combining both associative and categorical knowledge into a similarity measure between the current recalled entity and the next. The classic psychological experiment for this model is the Semantic (or, Verbal) Fluency test:

1. Recall and record an object (*e.g.*, an animal);
2. Record the next object of this type you think of;
3. Continue recording for the remaining time

The reader is encouraged to try this process out for themself. One advantage of this test lies in not restricting (or having

[1] Though less applicable in technical or domain-specific corpuses where examples are too few and far between, this is the intuition that leads to the success of vector-based semantic embeddings like `gloVe` or `word2vec` [23, 24].

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

to specify apriori) the relationship between objects required to record subsequent ones. For example:

$$\text{dog} \rightarrow \text{cat} \rightarrow \text{tiger} \rightarrow \text{lion} \rightarrow \text{elephant} \rightarrow \text{wolf} \cdots$$

For example, it is common for animal-based semantic fluency lists to start with household pets, potentially switching to entirely unrelated categories like "large cats," for further exploration, before either retracing back to a previous category (*e.g.*, canines to "wolf" via "dog") or onward via new similarities (*e.g.*, African animals to "elephant" via "lion"). Different people can create different fluency lists, owing to differences in how they psychologically structure relationships between concepts.[2]

The key contribution of this work is to propose that explicitly modeling this process lends itself well to recovering engineering knowledge from text-based technical records. While, technicians are not purely sampling arbitrary system concepts, as you might a list of animals, we nevertheless assume that each subsequent concept written in a MWO is directly conditional on what was written previously. [3] Then, an MWO consists of "jumps" between concepts that depend upon previously "visited" concepts. This assumption allows us to infer relationships between concepts given examples of MWOs. This boils down to two key components of the technician's cognitive task when recalling relevant information to write down MWO's:

- A technician records concepts sequentially, as he or she recalls unique defining characteristics of the MWO.
- They recall these characteristics by remembering links between them, and any recently recalled characteristics.

This differs from a standard Bag of Words model—where all entities occurring in a document are assumed to be linked through co-occurrence—and from n^{th}-order language models—where relations are limited to the nearest (or, previous) n entities. In technical shorthand (like MWOs), objects listed later on may be linked to any of the previously mentioned objects, not strictly those directly adjacent to it. For instance, the MWO *"Leaking hydraulic valve; cleaned oil spill and replaced O-ring"* consists of a sequence of concepts (*leak→hydraulic→valve...*), not all of which share the same causal structure: perhaps "hydraulic", "valve", and "leak" are all potentially subsets of a hydraulic "system", but "replace", "clean", and "oil" all have potential to span subsystems. Similarly, in this MWO, "oil" would likely be considered as linked with "leak", more than it would to the *closer* entity "replace". This illustrates nicely the trade-off between categorical and associative memory foraging that [36] discusses at length, and is precisely the feature of MWOs we exploit to extract a more sparse representation of system relationships through the Initial-Visit Emitting Random Walks semantic fluency model, which we detail next.

3.2 Initial-Visit Emitting Random Walks

Based on the above modelling assumptions, we demonstrate the application of an Initial-Visit Emitting Random Walks (INVITE) as initially described by [34], on recovering system structures from MWOs. Say the set of components or concepts in our system is denoted by the node-set n. A set of T tags [4] can be denoted as a Random Walk (RW) trajectory $\mathbf{t} = \{t_1, t_2, t_3, \cdots t_T\}$, where $T \leq n$. However, this limit on the size of T assumes tags are a set of unique entries: any transitions between previously visited nodes in $\mathbf{t}$ will not be directly observed, making the transitions observed in $\mathbf{t}$ strictly non-Markovian, and allowing for a potentially infinite number of possible paths to arrive at the next tag.

Instead of directly computing over this intractable model for generating $\mathbf{t}$, the key insight from the original INVITE paper comes from partitioning $\mathbf{t}$ into $T-1$ Markov chains with absorbing states, where previously visited nodes are transient states, and unseen nodes are absorbing. It is then possible to calculate the absorption probability into the k^{th} transition ($t_k \rightarrow t_{k+1}$) using the *fundamental matrix* of each partition. If the partitions at this jump consist of q transient states with transition matrix amongst themselves $\mathbf{Q}^{(k)}_{q\times q}$, and r absorbing states with transitions into them from q as $\mathbf{R}^{(k)}_{q\times r}$, the Markov chain $\mathbf{M}^{(k)}_{n\times n}$ has the form

$$\mathbf{M}^{(k)} = \begin{pmatrix} \mathbf{Q}^{(k)} & \mathbf{R}^{(k)} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{2}$$

where $\mathbf{0}$, $\mathbf{I}$ represent lack of transition between/from absorbing states. It follows from [38] that the probability P of a chain starting at t_k being absorbed into state $k+1$, letting $\mathbf{N} = (\mathbf{I}-\mathbf{Q})^{-1}$, is given as

$$P(t_{k+1}|t_{1:k}, \mathbf{M}) = \mathbf{N}^{(k)} R^{(k)}\Big|_{q,1} \tag{3}$$

The probability of being absorbed at $k+1$ conditioned on jumps $1:k$ is thus equivalent to the probability of observing the

[2] *e.g.*

$$\text{dog} \rightarrow \text{walk} \rightarrow \text{run} \rightarrow \text{gym} \rightarrow \cdots \quad \text{vs}$$
$$\text{dog} \rightarrow \text{home} \rightarrow \text{family} \rightarrow \text{meal} \rightarrow \cdots$$

[3] This is standard practice in the language modeling domain [37].

[4] While traditional application of "tagging" assumes the set of labels to be strictly un-ordered (as in multi-label classification), we follow [15,35] by assuming tags are generated directly from text by keyword recognition. It is thereby trivial to reverse the process, assigning each tag a position as the first time its corresponding keyword was recognized in the original text.

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

$k+1$ INVITE tag. If we approximate an a priori distribution of tag probabilities to initialize our chain as $t_1 \sim \text{Cat}(n, \theta)$ (which could be empirically derived or simulated), then the likelihood of our observed tag chain $\mathbf{t}$, given a Markov chain, is

$$\mathscr{L}(\mathbf{t} \,|\, \theta\,; \mathbf{M}) = \theta(t_1) \prod_{k=1}^{T-1} P(t_{k+1} \,|\, t_{1:k}\,; \mathbf{M}) \tag{4}$$

Finally, if we observe a corpus of tag lists $\mathbf{C} = \{\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_c\}$, and assume θ can be estimated separately from $\mathbf{M}$, then we can finally frame the problem as minimizing our loss function, the negative log-likelihood of our corpus over $\mathbf{M}$:

$$\mathbf{M}^* = \underset{\mathbf{M}}{\arg\min} \quad \sum_{i=1}^{C} \sum_{k=1}^{T_i-1} -\log P\left(t_{k+1}^{(i)} \middle| t_{1:k}^{(i)}, \mathbf{M}\right) \tag{5}$$

3.3 Implementation

As stated in Eq. 5, the optimization is constrained; in addition to requiring row-stochasticity, the matrix N is only guaranteed to exist if self-transitions are disallowed, as proved in [34]. Similar to that implementation, we introduce a softmax re-parameterization of $\mathbf{M}$ that allows the optimization to be unconstrained in $\mathbb{R}^{n \times n}$, and guaranteeing row-stochasticity.

$$M_{i,j} \leftarrow \frac{\exp(M_{i,j})}{\left[\sum_j \exp(\mathbf{M}_i)\right]_j}$$

However, we introduce several modifications to this re-parameterization:

Edge Weights Because it is important for our purposes to estimate the weight (*i.e.*, importance) of each relationship, we to not require (as in [39]) that the structure of $\mathbf{M}$ is un-weighted—in this case each relationship would either exists or not exist. However, sparsity of $\mathbf{M}$ is still desirable, so we apply an L_1-penalty to the loss function, adding an $(\alpha/T) \cdot \|\mathbf{M}\|_1$ term to Eq. 5. The parameter α should generally be tuned via cross-validation where possible, but to demonstrate effectiveness in an unsupervised setting (as is expected to be the case when no "true" $\mathbf{M}$ is yet known), we use $\alpha = 0.01$, which was found to be robust to sensitivity trials for one log-factor in either direction.

Edge Direction In addition, Eq. 5 implies that $\mathbf{M}$ represents a *directed* graph. Though we model each tag as being generated conditional on preceding tags alone, we wish to preserve the intuition that relationships between tags are still assumed to be bi-directional, while not strictly enforcing $\mathbf{M}$ to be symmetric (undirected), as in [39]. Put simply, one-directional relationships can be useful when they are strictly the case (*e.g.*, oil→leak), but we may not wish to encourage one-directional relations that are quirks of imbalanced data and how people talk (gear 1 ↔ gear 2) To ensure the recovered weights in each direction are meaningful, and to speed-up recovery of what we assume is a "symmetry-dominant" $\mathbf{M}$, we bias it toward symmetry via an update to each entry prior to softmax:

$$M_{i,j} \leftarrow \max\left\{M_{i,j}, M_{j,i}\right\}$$

Because of these alterations, the analytic gradient for the INVITE loss function described in [34] no longer applies; instead, we make use of automatic differentiation as a means to ensure accurate gradient calculations under the above modifications [40]. The package `autograd` [41] was used to exploit a number of convenience functions for doing so, in the Python programming language.

4 EXPERIMENTS

The first experiment demonstrates the tractability of the INVITE model in the context of MWO-type data by generating synthetic MWOs from real engineering systems as described in [42]. We use these synthetic MWOs to (1) measure the network recovery accuracy of the INVITE model, (2) compute the sample efficiency of the INVITE model, and (3) compare the INVITE model to co-occurrence similarity thresholding models currently used in the state-of-the-art.

Second, we apply our proposed method to a corpus of real excavator MWOs, for which a "true" underlying structure does not yet exist. We compare the plausibility of work orders sampled from our network estimation to the original dataset and benchmark our model with respect to purely associative sampling.

Due to the high dimensionality of Eq. 5, and the noisy nature of observations, we use Stochastic Gradient Descent (SGD) to perform optimization of $\mathbf{M}$. Specifically, we use the ADAM algorithm [43], which modifies the gradient estimation for each iteration with first- and second-order momentum estimates from previous iterations, improving convergence behavior. Because each tag transition is considered a reliable observation, and the underlying structure of $\mathbf{M}$ is generally sparse relative to the complete adjacency graph between the set of all tags, a learning rate of 0.9 was used, but with minibatches of 5 censored lists each. Exponentially-weighted learning-rate decay was used, along with time-discounted averaging of $\mathbf{M}$, with settings as suggested by [44].

4.1 Exp. 1: Recovering Known Engineering Networks

To validate the ability of our method to accurately reconstruct engineering networks under varying data quantities, we

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

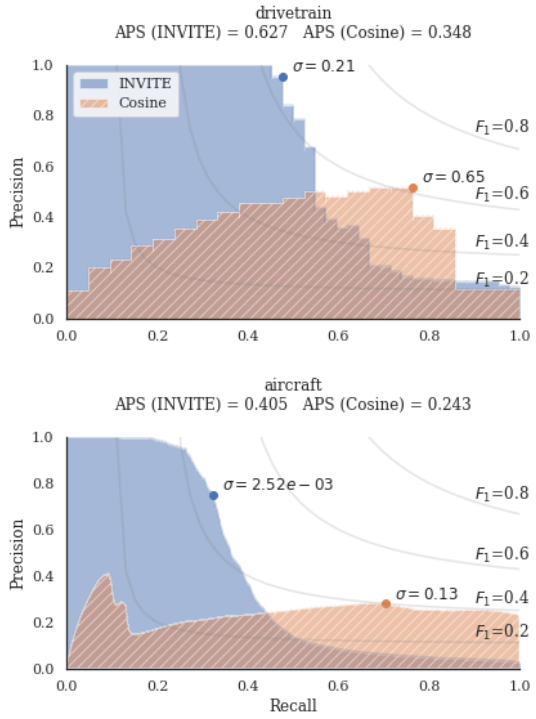


FIGURE 1: Comparing INVITE and Cosine-Similarity thresholding performance for recovering true network structure. **Top**: Recovery performance (precision vs. recall) for the *drivetrain* network. Trained on 18 samples ("work orders"), at 3 tags each. **Bottom**: Same comparison for the more complex *aircraft* network, trained on 1634 samples at 5 tags each. Also shown are the F_1-score iso-lines, along with F_1-optimal thresholds (σ) for each model setting.

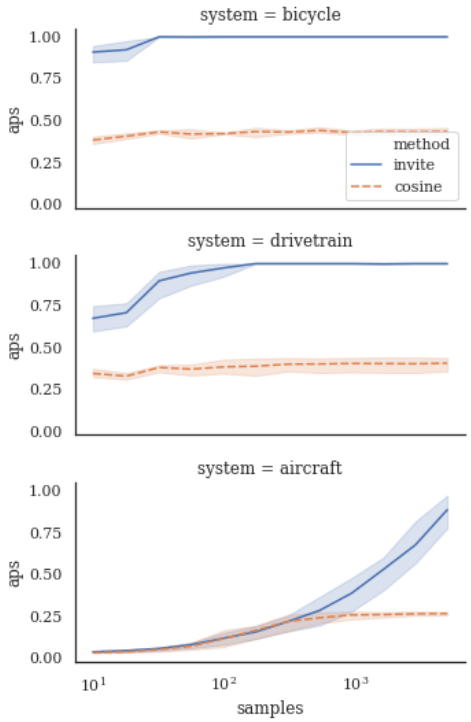


FIGURE 2: Mean average precision score (APS) for the three system networks of [42], shown with mean APS over sample lengths $T \in \{3,4,5\}$, and a 1000-bootstrap-sample 95% confidence interval. INVITE consistently outperforms similarity thresholding in low-data, low-complexity scenarios. In complex networks, performance is comparable until a significant number of samples are available, after which a lack of sparsity causes the cosine method's performance to plateau.

first synthesize censored tag lists from true component networks described in [42, 46]: a bicycle ($n = 10$), an automotive drivetrain ($n = 18$), and an aircraft ($n = 375$). Drawn layouts for each network are provided for reference in the appendix. For each network, censored random walks were generated by performing a random walk over the nodes until either 100 transitions or all nodes have been visited. The first unique visit to each node was recorded to simulate censoring, and the lists were clipped to the first 3,4, or 5 node visits, to reflect the typical number of tags seen in real MWO datasets (see Exp. 2, below, for an example). The number of censored lists used to train the models was evenly sampled at 11 intervals on a log-scale from 10 - 5000 lists, for a total $3 \times 11 \times 3 = 108$ trials.

Because the original networks are relatively sparse (See Table 1), the classification of edges as "existing" or "not existing" can be framed as a class-imbalanced information retrieval problem. Given some measure of node similarities (entries in the recovered adjacency matrix), we wish to threshold $\mathbf{M}$ such that, for a given threshold value $\sigma \in [0,1]$, the entries of a thresholded adjacency matrix $\mathbf{M}^\sigma$ are given by:

$$M^\sigma_{i,j} = \begin{cases} 1, & \text{if } M^*_{i,j} \geq \sigma \\ 0, & \text{otherwise} \end{cases}$$

Prior to selecting a specific threshold, it is useful to recognize how robustly each model performs under varying threshold values, since the underlying "true" networks are available. For class-imbalanced learning problems like this, precision-recall (*P-R*) curves can elucidate model robustness under varying threshold sensitivities [47]. In Fig. 1, the precision-recall curves for

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

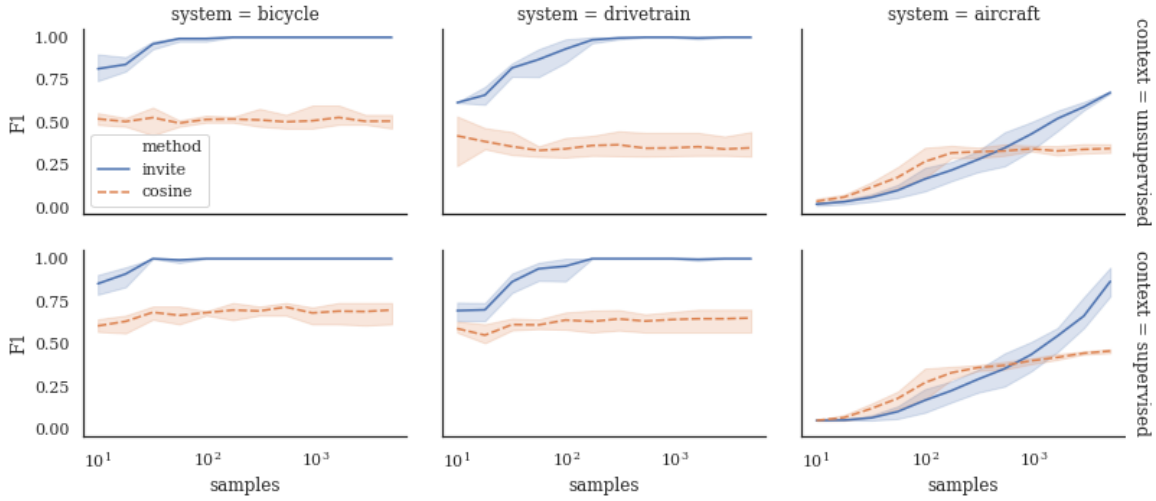


FIGURE 3: Mean F_1 reconstruction scores for the three system networks of [42], shown with mean score over sample lengths $T \in \{3,4,5\}$, and a 1000-bootstrap-sample 95% confidence interval. In an unsupervised context (top row), thresholds for node similarity were selected using a knee-finding heuristic [45], for where the EDF of edge-weights showed maximum curvature. In a supervised context (bottom row), the optimal threshold was selected as one that maximized the model's F_1-score. The INVITE method significantly outperforms pure co-occurrence similarity thresholds as the number of samples increases, and because the EDF is much more spread-out for cosine similarities, picking a "good" threshold is much more difficult than the sparsity-inducing INVITE models.

TABLE 1: Engineering component network summary for Experiment (1). Network models adapted from [42].

Model	Nodes	Sparsity
bicycle	10	80.0%
drivetrain	18	88.4%
aircraft	375	97.5%

the drivetrain model demonstrate that INVITE can quickly recover simpler networks with under 20 observations at relatively few "tags" each, while Cosine Similarity only robustly captures the global structure of the network—precision is relatively invariant over wide ranges of recall. This is even more pronounced for more complex networks, with the INVITE model capable of achieving either high precision or high recall, while the Cosine Similarity threshold has difficulty improving it under any circumstance. One way to summarize this robustness under varying threshold is calculate the average the precision score (APS) gained by each threshold's increase of recall R:

$$\text{APS} = \sum_{\sigma} [R(\sigma_i) - R(\sigma_{i-1})] P(\sigma_i) \tag{6}$$

The APS score will not give a "good" σ, but instead summarizes the total "goodness" of each model across possible σ. APS scores for the INVITE and Cosine Similarity models are shown against training set size in Fig. 2. APS eventually plateaus for the cosine model in every case. INVITE can perfectly recover the bicycle and drivetrain structures after around 100 samples. For the aircraft network, while INVITE has nearly identical performance to cosine similarity below 500 samples, INVITE's APS almost reaches 1.0 with 5000 samples.

In practice, selecting the value for σ will depend on whether training examples are available: if not, a heuristic threshold such as knee-finding can be applied; if examples are available, it is possible to use performance measures appropriate for imbalanced learning problems (*e.g.*the F_1-score), and optimize the threshold for this value. In the common case that no training labels are available (no "true" structures are known), a common heuristic for thresholding values posits that diminishing returns occur for the retrieval function after the point of maximum curvature on the empirical distribution function (EDF) of values to threshold at the point of diminishing returns—*e.g.*, using a so-called "knee-finding" algorithm. To test the performance of both the cosine-similarity (bag-of-words) and INVITE recovered networks with respect to the originals, we apply the `kneedle` algorithm [45] to calculate a threshold σ. The F_1-score can then be calculated for $\mathbf{M}^{\sigma}$ as for each training-set size (see unsuper-

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

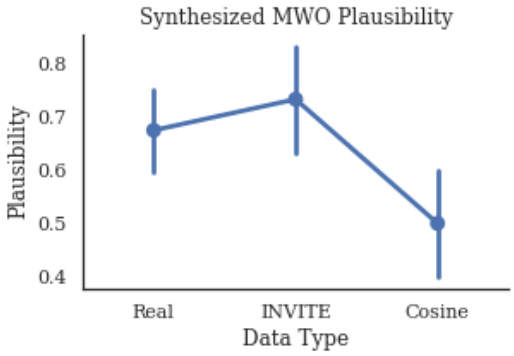


FIGURE 4: Mean plausibility ratio for 100 MWOs, both real and synthesized by sampling $\mathbf{M}^{\sigma}$ recovered from INVITE and Cosine methods. Confidence intervals show the inter-quartile range of 1000 bootstrap samples.

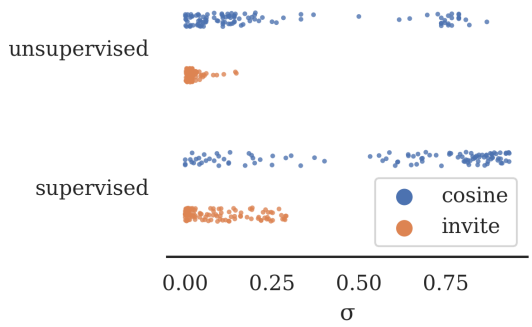


FIGURE 5: Thresholds selected by knee-finding heuristic (unsupervised) and optimal F_1-score (supervised). The cosine similarity performance was more sensitive to σ selection than INVITE.

vised context in the top row of Fig. 3). If parts of the underlying structure are known a priori, it is possible to tune σ so that the F_1-score of $\mathbf{M}^{\sigma}$ vs. the "test-set" (the true $\mathbf{M}$) is maximized. These can likewise be found on the bottom of Fig. 3

4.2 Exp. 2: Real-World Excavator MWOs

To assess the applicability of INVITE to real-world scenarios, we apply our model to tags annotated for a mining dataset (8264 MWOs) pertaining to 8 similarly-sized excavators at various sites across Australia [15, 48]. The tags were created by a subject-matter expert spending 1 hour of time in the annotation assistance tool `nestor` [49], using a methodology outlined in [50]. The tag annotations were limited to objects (bolt, motor, fan, *etc.*), problems (leak, missing, cracked, *etc.*), and solutions (replace, repair, stick, *etc.*) that occurred at least 50 times each in the original corpus, for a total of 77 unique tags. Subsequently, the same settings for solving Eq. 5 were used as in the previous experiment, though the optimization was initialized with the cosine similarity matrix to speed convergence.

To test whether the INVITE model was able to learn a robust representation of the system structure, we perform blind tests of the generative capability of each recovered network. First, the starting tag probability θ was set as the observed distribution of first tags in the original dataset. Then, censored random walks of length $T = 5$ were sampled from both an INVITE and a cosine-similarity recovered network, without thresholding. This is intented to preserve weighted relationships between tags, for the purposes of data synthesis. The expert was then given a list of 100 randomly mixed MWOs, made from 40 real work-orders,[5] 30 INVITE censored lists, and 30 cosine-similarity censored lists. The resulting lists were filtered to only contain lists of tags not explicitly found in the original data. The expert was then asked to blindly classify each MWO as being "plausible" or "not plausible", such that an MWO would or would not reasonably occur based only on the tags in each. The fraction of real, INVITE, and cosine-generated work orders marked as "plausible" can be found in Fig. 4. Both the real and INVITE-synthesized MWOs are within a similar plausibility range, between $60\% - 80\%$, while the cosine similarity MWOs are between $40\% - 60\%$ plausible, overall.

5 DISCUSSION

To enable optimization in continuous space, our model does not enforce un-weighted, un-diriected graphs, as in Zemla & Austerweil's extension of the INVITE model that optimizes in discrete space [39]. Their version is intended to induce sparsity without artificially introducing new tuning parameters, as we have in introducing σ and α. In practice, however, it is reasonable to select a low-valued positive $\sigma \ll 1$, or use a knee-finding heuristic on the EDF of edge weights. This is because the L_1 penalty on edge weights, plus the tendency of INVITE to route random walks through commonly-visited chokepoint nodes, naturally drives "unnecessary" edge weights to near-zero probability. As seen in Fig. 5, the "knee" in the edge-weight EDF for INVITE was nearly always near-zero, for all experiments. Even when oracle information was allowed to tune σ, the best F_1-score was still to be found strictly below $\sigma = 0.3$, with a majority below 0.1. In this sense, the signal-to-noise ratio of our INVITE model is quite high, making the selection of a good σ much less

[5] The process in [49] displays extracted concepts in order of their statistical importance, for the purposes of keyword recognition. As such, the annotator does not interact with the original work-orders directly during tagging.

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

difficult.

In truth, this thresholding does not completely solve the problem of knowledge extraction. If the goal of automatically extracting knowledge graphs is to suggest whether causal relationships *exist* between tags, and not primarily to synthesize data, weights are not needed in communicating these links, and may obfuscate important understanding below vastly more obvious relationships.

Another issue related to thresholding is how the row-stochastic constraint affects edge-value distribution: technically, each row in our model will be re-normalized every iteration, independent of other rows. This means that the row-normalization inherently de-symmetrizes $\mathbf{M}$. In reality, though we might model some asymmetry in node relations[6], the modality of direction being discovered by sentence-structure (ordering of the written tags) is not equivalent to the types of directionality we might want to discover. In memory, it could be beneficial to assume that the probability of transitioning between tags should be bi-directional, and allow desirable directionality to be proxied by local tree-like structures that reduce centrality of tags farther out. This bi-directional assumption implies making $\mathbf{M}$ a doubly-stochastic matrix. This has the added benefit of placing $\mathbf{M}$ on a simplex, *i.e.*, the space of permutation-invariant matrices belonging to the Birkhoff Polytope. There are recent developments [51, 52] in this space that could prove highly useful at reducing the state-space we search over.

Finally, our method is not intended to serve as a complete, end-to-end processing of natural language text into structured knowledge. Ultimately, the final structuring will need to be performed by humans. Instead, we believe the most efficient tools to assist in knowledge recovery will pose annotation questions in a lower-dimensional state-space, easier for a human to verify or edit *quickly*. Figure 6 illustrates how we believe INVITE takes steps toward this goal. Common structure recovery techniques, like cosine-similarity thresholding, tend to discover global structure quite well, but over-estimate the connectivity of local communities where hierarchical relationships are unknown-yet-assumed by the data. By definition, co-occurrence (bag-of-words) metrics are treating these work-orders more like un-censored random walks, starting from any tag and transitioning to any other in the list. Consequently, the local resolution of the structure it approximates is going to be fundamentally limited for tree-like communities, more reflecting a 2^{nd}- or 3^{rd}-order power graph[7] of the true, underlying structure.

In contrast, INVITE tends to concentrate edges to nodes that are highly central, forming "chokepoints" where global-scale transition mechanisms are unknown, but largely preserving local tree-like structures in outer communities. From an active learning perspective, humans tend to be quite good at verifying global-scale connections as viable or not—editing spurious connections in every over-dense local community is a much more difficult task for us than recognizing spurious individual connections to a small set of highly abstract concepts.

We believe this feature can be exploited to create better knowledge-structuring assistance tools in an active learning context. Such a tool could additionally benefit from a recent explosion in interest for preserving hierarchical relationships in vector space, *e.g.*, via Poincaré embeddings [54]. Additional care must be taken to allow flexible annotation of different kinds of relationships,[8] and allow for multiple (potentially disagreeing) annotators, subsequently suggesting relationship types for review. We envision a type of "topic model" over the space of knowledge graphs [55], or potentially a set of independent "graph components" that maximally explain the distribution of edge types in a community [56].

6 CONCLUSIONS AND FUTURE WORK

This paper presented a method to recover a structured representation of engineering knowledge from unstructured written documents (specifically, Manufacturing Work Orders), based on initial-visit emitting random walks (INVITE). Compared to previous methods, our technique preserves local connectivity structures, even in tree-like communities. This can lead to (1) better generative capability for synthesizing plausible documents (such as work-orders) in a simulation context; and (2) allowing us to cast the knowledge-structuring problem in probabilistic context that is potentially amendable to active-learning; this can minimize the number of local-scale edits needed relative to global-scale, abstract connections that humans can easily spot and correct.

Overall, the model we describe here can enable experts and novices alike to benefit from tacit system knowledge contained within frequently unused mountains of technical work-orders, by quickly prototyping computable representations of this knowledge for downstream usage in analysis pipelines. We believe that by explicitly incorporating cognitive theories into our modeling assumptions about how technicians might represent and then recall their knowledge in maintenance work-orders, we can accelerate the training and use of unsupervised data-driven expert systems in engineering design.

ACKNOWLEDGMENT

Thanks to Dr. Michael Brundage (NIST) for his efforts annotating and rating MWOs, and to Dr. Melinda Hodkiewicz (Univ. Western Australia) for providing the excavator data, and for many enjoyable and enlightening discussions on this topic.

[6] *e.g.*in hierarchies: "gear-1" may be a member of "gearbox," making the link gear-1 $\rightarrow$ gearbox a stronger link in a technician's head than the other direction.

[7] The graph G's n^{th}-order powergraph $P(G,n)$ has an edge between any two nodes if the minimum path length between those nodes in G is at most n.

[8] *e.g.*, Walsh *et al.*actually construct three types of structured system representations: functional, parametric, and component (which we use here)

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

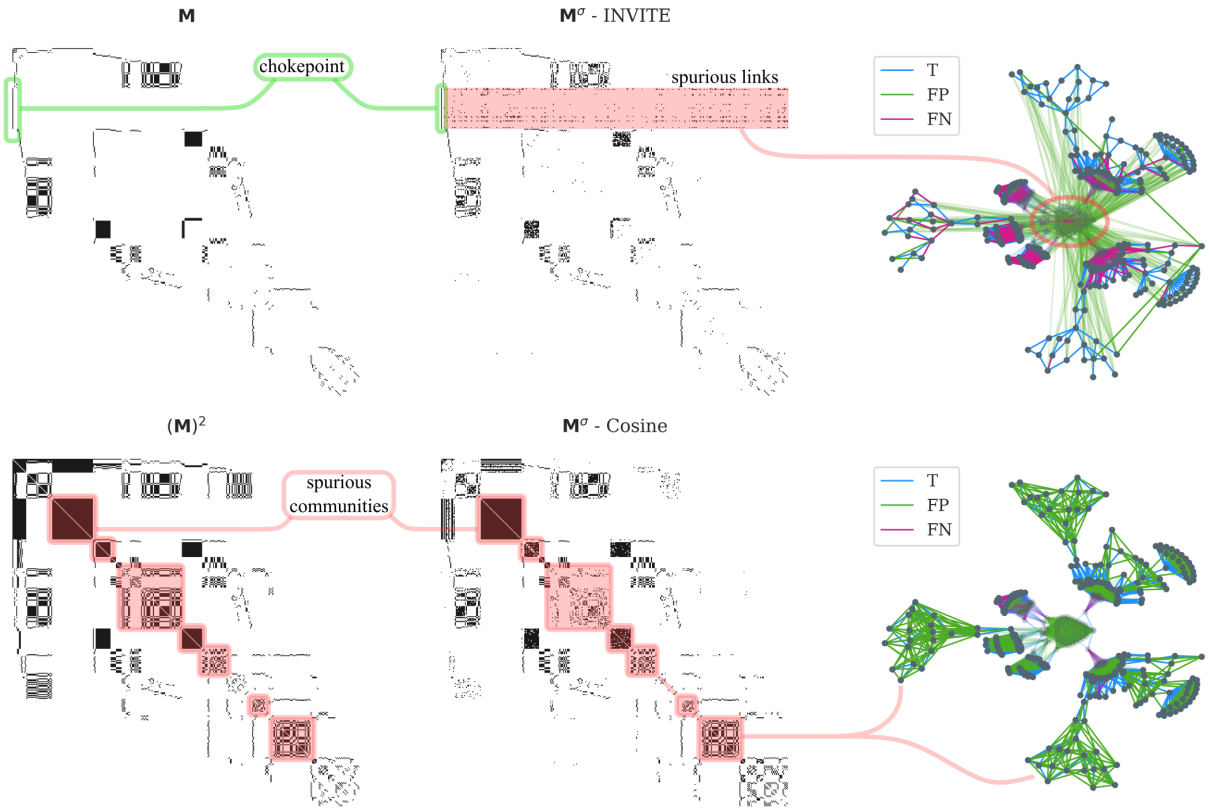


FIGURE 6: Comparing aircraft model network reconstruction for F_1-optimal INVITE ($F_1 = 0.75$) and Cosine Similarity ($F_1 = 0.49$) methods. Shown are the original "true" adjacency matrix $\mathbf{M}$, its 2^{nd}-order power graph $(\mathbf{M})^2$, and the thresholded adjacency matrices $\mathbf{M}^\sigma$ for both INVITE and Cosine Similarity. For visualization, the matrix rows/columns are sorted by the closeness centrality of each node [53], better indicating which nodes form core/integral components in the system (upper-left) and which are more likely a part of localized "edge" communities (bottom and right). These edge communities are highlighted in the graph layouts on the right, where nodes in the top 25^{th} percent most central (and their edges) are transparent. INVITE directly estimates the underlying structure of $\mathbf{M}$ by accounting for node censoring in observations, concentrating uncertain edges into a few highly-connected "chokepoint" nodes. Cosine similarity mistakes co-occurrence of components in a sample for direct relationships, forming dense, spurious communities throughout the graph that are reflective of higher-order powers of $\mathbf{M}$, as shown here. Concentrating the false-positive relationships (FP) in a few highly central nodes makes INVITE a viable candidate for querying human experts for annotation/critique in an active-learning context.

7 DISCLAIMER

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

REFERENCES

[1] Guimerà, R., and Sales-Pardo, M., 2009. "Missing and spurious interactions and the reconstruction of complex networks". *Proceedings of the National Academy of Sciences,* ***106***(52), pp. 22073–22078.

[2] Gomez-Rodriguez, M., Leskovec, J., and Krause, A., 2012. "Inferring networks of diffusion and influence". *ACM Transactions on Knowledge Discovery from Data (TKDD),* ***5***(4), p. 21.

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

[3] Linderman, S., and Adams, R., 2014. "Discovering latent network structure in point process data". In International Conference on Machine Learning, pp. 1413–1421.

[4] De Paula, Á., Rasul, I., and Souza, P., 2018. "Recovering social networks from panel data: identification, simulations and an application".

[5] Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al., 2018. "Relational inductive biases, deep learning, and graph networks". *arXiv preprint arXiv:1806.01261*.

[6] Raissi, M., Perdikaris, P., and Karniadakis, G. E., 2017. "Machine learning of linear differential equations using gaussian processes". *Journal of Computational Physics,* ***348***, pp. 683–693.

[7] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al., 2017. "Visual genome: Connecting language and vision using crowdsourced dense image annotations". *International Journal of Computer Vision,* ***123***(1), pp. 32–73.

[8] Speer, R., Chin, J., and Havasi, C., 2017. "Conceptnet 5.5: An open multilingual graph of general knowledge". In Thirty-First AAAI Conference on Artificial Intelligence.

[9] ISO 15926-1:2004, 2004. Industrial automation systems and integration – Integration of life-cycle data for process plants including oil and gas production facilities – Part 1: Overview and fundamental principles. Standard, International Organization for Standardization, Geneva, CH, July.

[10] Leal, D., 2005. "ISO 15926" life cycle data for process plant": An overview". *Oil & gas science and technology,* ***60***(4), pp. 629–637.

[11] ISO/TS 15926-8:2011, 2011. Industrial automation systems and integration – Integration of life-cycle data for process plants including oil and gas production facilities – Part 8: Implementation methods for the integration of distributed systems: Web Ontology Language (OWL) implementation. Standard, International Organization for Standardization, Geneva, CH, Oct.

[12] Batres, R., West, M., Leal, D., Price, D., Masaki, K., Shimada, Y., Fuchino, T., and Naka, Y., 2007. "An upper ontology based on ISO 15926". *Computers & Chemical Engineering,* ***31***(5-6), pp. 519–534.

[13] Klüwer, J. W., Skjæveland, M. G., and Valen-Sendstad, M., 2008. "Iso 15926 templates and the semantic web". In Position paper for W3C Workshop on Semantic Web in Energy Industries; Part I: Oil and Gas.

[14] Kumar, N., Kumar, M., and Singh, M., 2016. "Automated ontology generation from a plain text using statistical and nlp techniques". *International Journal of System Assurance Engineering and Management,* ***7***(1), pp. 282–293.

[15] Hodkiewicz, M., and Ho, M. T.-W., 2016. "Cleaning historical maintenance work order data for reliability analysis". *Journal of Quality in Maintenance Engineering,* ***22***(2), pp. 146–163.

[16] Ho, M., 2015. "A shared reliability database for mobile mining equipment". PhD thesis, University of Western Australia.

[17] Eppinger, S. D., and Browning, T. R., 2012. *Design structure matrix methods and applications*. MIT press.

[18] Browning, T. R., 2016. "Design structure matrix extensions and innovations: a survey and new opportunities". *IEEE Transactions on Engineering Management,* ***63***(1), pp. 27–52.

[19] Ellinas, C., Allan, N., Durugbo, C., and Johansson, A., 2015. "How robust is your project? from local failures to global catastrophes: A complex networks approach to project systemic risk". *PloS one,* ***10***(11), p. e0142469.

[20] Robertson, S., 2004. "Understanding inverse document frequency: on theoretical arguments for idf". *Journal of documentation,* ***60***(5), pp. 503–520.

[21] Steyvers, M., and Griffiths, T., 2007. "Probabilistic topic models". *Handbook of latent semantic analysis,* ***427***(7), pp. 424–440.

[22] Blei, D. M., Griffiths, T. L., and Jordan, M. I., 2010. "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies". *Journal of the ACM (JACM),* ***57***(2), p. 7.

[23] Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013. "Efficient estimation of word representations in vector space". *arXiv preprint arXiv:1301.3781*.

[24] Pennington, J., Socher, R., and Manning, C., 2014. "Glove: Global vectors for word representation". In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.

[25] Sharp, M., Sexton, T., and Brundage, M. P., 2017. "Toward semi-autonomous information". In IFIP International Conference on Advances in Production Management Systems, Springer, pp. 425–432.

[26] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M., 2009. "Reading tea leaves: How humans interpret topic models". In Advances in neural information processing systems, pp. 288–296.

[27] Strohmaier, M., Körner, C., and Kern, R., 2012. "Understanding why users tag: A survey of tagging motivation literature and results from an empirical study". *Web Semantics: Science, Services and Agents on the World Wide Web,* ***17***, pp. 1–11.

[28] Vander Wal, T., 2007. Folksonomy. http://vanderwal.net/folksonomy.html.

[29] Specia, L., and Motta, E., 2007. "Integrating folksonomies with the semantic web". In European semantic web conference, Springer, pp. 624–639.

[30] Mousselly-Sergieh, H., Egyed-Zsigmond, E., Gianini, G.,

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

Döller, M., Kosch, H., and Pinon, J.-M., 2013. "Tag similarity in folksonomies". In INFORSID, Vol. 29, Inforsid, pp. 319–334.

[31] Heymann, P., and Garcia-Molina, H., 2006. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Tech. rep., Stanford.

[32] Henschel, A., Woon, W. L., Wachter, T., and Madnick, S., 2009. "Comparison of generality based algorithm variants for automatic taxonomy generation". In Innovations in Information Technology, 2009. IIT'09. International Conference on, IEEE, pp. 160–164.

[33] Schvaneveldt, R. W., Durso, F. T., and Dearholt, D. W., 1989. "Network structures in proximity data". In *Psychology of learning and motivation*, Vol. 24. Elsevier, pp. 249–284.

[34] Jun, K.-S., Zhu, J., Rogers, T. T., Yang, Z., et al., 2015. "Human memory search as initial-visit emitting random walk". In Advances in neural information processing systems, pp. 1072–1080.

[35] Sexton, T., Brundage, M. P., Hoffman, M., and Morris, K. C., 2017. "Hybrid datafication of maintenance logs from ai-assisted human tags". In 2017 IEEE International Conference on Big Data (Big Data), IEEE, pp. 1769–1777.

[36] Hills, T. T., Todd, P. M., and Jones, M. N., 2015. "Foraging in semantic fields: How we search through memory". *Topics in Cognitive Science,* ***7***(3), pp. 513–534.

[37] Lv, Y., and Zhai, C., 2009. "Positional language models for information retrieval". In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 299–306.

[38] Doyle, P. G., and Snell, J. L., 2000. "Random walks and electric networks". *arXiv preprint math/0001057*.

[39] Zemla, J. C., and Austerweil, J. L., 2018. "Estimating semantic networks of groups and individuals from fluency data". *Computational Brain & Behavior,* ***1***(1), pp. 36–58.

[40] Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M., 2018. "Automatic differentiation in machine learning: a survey". *Journal of Marchine Learning Research,* ***18***, pp. 1–43.

[41] Maclaurin, D., 2016. "Modeling, inference and optimization with composable differentiable procedures". PhD thesis.

[42] Walsh, H. S., Dong, A., and Tumer, I. Y., 2019. "An analysis of modularity as a design rule using network theory". *Journal of Mechanical Design,* ***141***(3), p. 031102.

[43] Kingma, D. P., and Ba, J., 2014. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980*.

[44] Bottou, L., 2012. "Stochastic gradient descent tricks". In *Neural networks: Tricks of the trade*. Springer, pp. 421–436.

[45] Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B., 2011. "Finding a" kneedle" in a haystack: Detecting knee points in system behavior". In 2011 31st International Conference on Distributed Computing Systems Workshops, IEEE, pp. 166–171.

[46] Haley, B. M., Dong, A., and Tumer, I. Y., 2016. "A comparison of network-based metrics of behavioral degradation in complex engineered systems". *Journal of Mechanical Design,* ***138***(12), p. 121405.

[47] Saito, T., and Rehmsmeier, M., 2015. "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets". *PloS one,* ***10***(3), p. e0118432. An optional note.

[48] Hodkiewicz, M. R., Batsioudis, Z., Radomiljac, T., and Ho, M. T., 2017. "Why autonomous assets are good for reliability–the impact of 'operator-related component'failures on heavy mobile equipment reliability". In Annual Conference of the Prognostics and Health Management Society 2017.

[49] Madhusudanan Navinchandran, F., Bones, L., Brundage, M., Hoffman, M., Moccozet, S., and Sexton, T., 2018. Nestor: a toolkit for quantifying tacit maintenance knowledge, for investigatory analysis in smart manufacturing.

[50] Sexton, T., Hodkiewicz, M., Brundage, M. P., and Smoker, T., 2018. "Benchmarking for keyword extraction methodologies in maintenance work orders". In PHM Society Conference, Vol. 10.

[51] Adams, R. P., and Zemel, R. S., 2011. "Ranking via sinkhorn propagation". *arXiv preprint arXiv:1106.1925*.

[52] Linderman, S. W., Mena, G. E., Cooper, H., Paninski, L., and Cunningham, J. P., 2017. "Reparameterizing the birkhoff polytope for variational permutation inference". *arXiv preprint arXiv:1710.09508*.

[53] Sabidussi, G., 1966. "The centrality index of a graph". *Psychometrika,* ***31***(4), pp. 581–603.

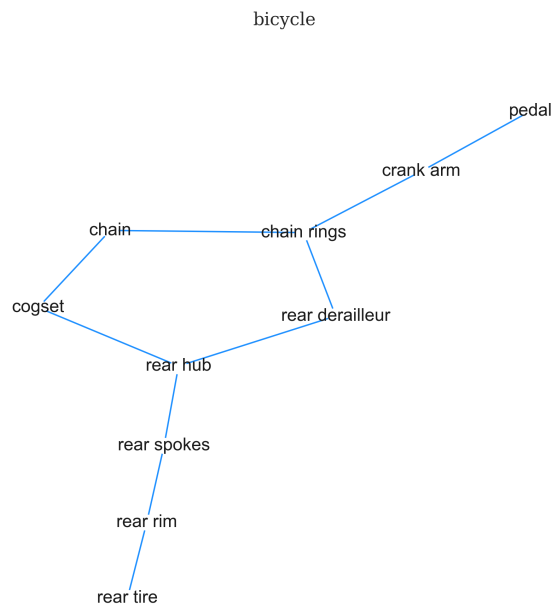
[54] Nickel, M., and Kiela, D., 2017. "Poincaré embeddings for learning hierarchical representations". In Advances in neural information processing systems, pp. 6338–6347.

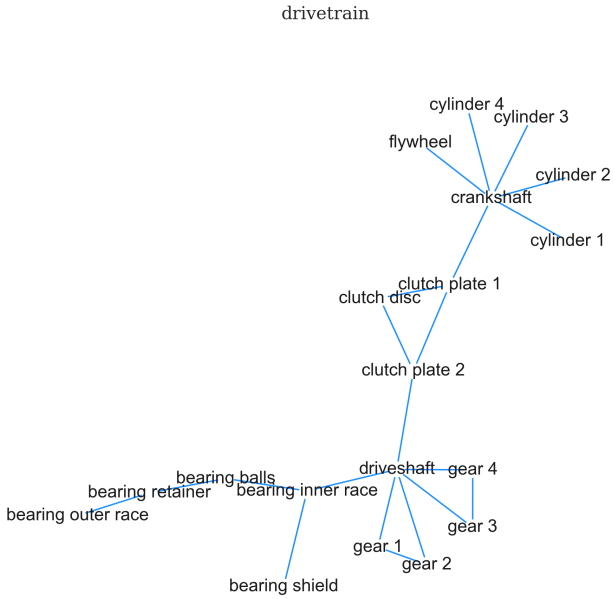
[55] Gerlach, M., Peixoto, T. P., and Altmann, E. G., 2018. "A network approach to topic models". *Science advances,* ***4***(7), p. eaaq1360.

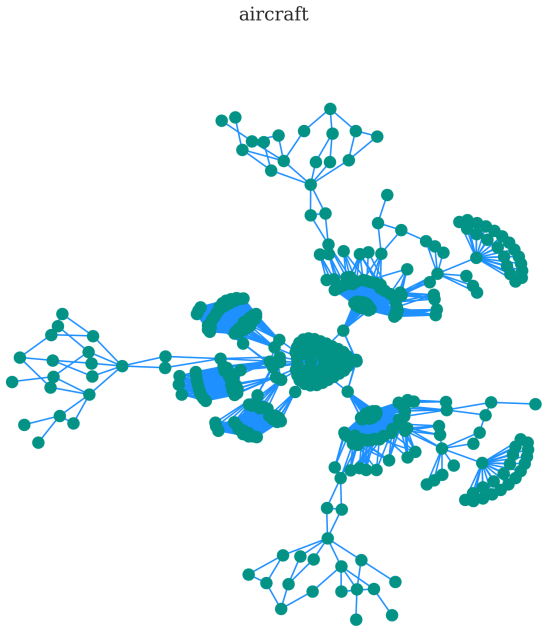
[56] Park, B., Kim, D.-S., and Park, H.-J., 2014. "Graph independent component analysis reveals repertoires of intrinsic network components in the human brain". *PloS one,* ***9***(1), p. e82873.

Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

A APPENDIX - WALSH *ET AL.* NETWORKS







Sexton, Thurston; Fuge, Mark.
"Using Semantic Fluency Models Improves Network Reconstruction Accuracy of Tacit Engineering Knowledge."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference. IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

Proceedings of the ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2019
August 18 - 21, 2019, Anaheim, California, USA

IDETC2019-97095

A STANDARDIZED REPRESENTATION OF CONVOLUTIONAL NEURAL NETWORKS FOR RELIABLE DEPLOYMENT OF MACHINE LEARNING MODELS IN THE MANUFACTURING INDUSTRY

Max Ferguson
Seongwoon Jeong
Kincho H. Law
Engineering Informatics Group
Stanford University
Stanford, California, 94305

Svetlana Levitan
IBM Cognitive Applications
IBM
Chicago, Illinois, 60606

Anantha Narayanan
Viterbi School of Engineering
University of Southern California
Los Angeles, California, 90007

Rainer Burkhardt
Artificial Intelligence R&D
SoftwareAG
San Diego, California, 92130

Tridivesh Jena
Artificial Intelligence R&D
Zementis*
San Diego, California, 92130

Yung-Tsun Tina Lee
Systems Integration Division
National Institute of Standards and Technology
Gaithersburg, Maryland, 20899

ABSTRACT

The use of deep convolutional neural networks is becoming increasingly popular in the engineering and manufacturing sectors. However, managing the distribution of trained models is still a difficult task, partially due to the limitations of standardized methods for neural network representation. This paper seeks to address this issue by proposing a standardized format for convolutional neural networks, based on the Predictive Model Markup Language (PMML). A number of pretrained ImageNet models are converted to the proposed PMML format to demonstrate the flexibility and utility of this format. These models are then fine-tuned to detect casting defects in Xray images. Finally, a scoring engine is developed to evaluate new input images against models in the proposed format. The utility of the proposed format and scoring engine is demonstrated by benchmarking the performance of the defect-detection models on a range of different computation platforms. The scoring engine and trained models are made available at `https://github.com/maxkferg/python-pmml`

INTRODUCTION

Convolutional neural networks (CNNs) are finding numerous real-world use cases in the manufacturing domain [1]. Recent research has demonstrated that convolutional neural networks can obtain state-of-the-art performance on tasks like casting defect detection [2], anomaly detection in fibrous materials [3], and classification of waste recycling [4]. Recent progress in transfer learning has greatly reduced training dataset requirements, allowing powerful models to be trained with relatively small datasets [2]. However, sharing and deploying trained models still remains a difficult and error-prone task. Modern CNNs often have hundreds of layers and millions of parameters, making the distribution and deployment of these models a challenging task. In current practice, models are often saved using a serialization format specific to the machine learning framework used to train the model. In most cases, this method provides a reliable method of saving trained models, but it greatly hinders interoperability between different machine learning frameworks. In this paper, we seek to address this issue by developing a standardized representation of CNNs, based on the Predictive Model Markup

*Zementis was acquired by SoftwareAG on 1/1/2017 .

Ferguson, Max; Jeong, Seongwoon; Law, Kincho; Narayanan, Anantha Narayanan; Levitan, Svetlana; Tridivesh, Jena; Lee, Yung-Tsun.
"A Standardized Representation of Convolutional Neural Networks for Reliable Deployment of Machine Learning Models in the Manufacturing Industry."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

Language (PMML).

With the development and adoption of the Predictive Model Markup Language, it is now much easier to train and evaluate predictive models on separate computers. PMML is an XML-based language that enables the definition and sharing of predictive models between applications [5]. It provides a clean and standardized interface between the software tools that produce predictive models, such as statistical or data mining systems, and the consumers of models, such as applications that depend upon embedded analytics [6]. With PMML it is easy to train a model with a statistical package such as R, and save the model in a standardized format for use in a real-world application.

For deployment, predictive models are normally evaluated by a scoring engine. A scoring engine is a piece of software specifically designed to load a model in a standardized format, and use it to evaluate new observations or data points. Scoring engines are responsible for executing the mathematical operations that transform model inputs into model outputs. To promote interoperability, scoring engines are normally written in languages such as Python, C++ or Java, which are supported by most embedded systems and computing environments. Therefore, it is feasible to run the same scoring engine on a development computer, a cloud server, or an embedded device, without modifying the scoring engine code. The development of standards-compliant scoring engines allows PMML models to be reliably evaluated on a range of devices.

Neural Networks: An artificial neural network (ANN) is an interconnected group of nodes, akin to the vast network of neurons in a brain. The ANN itself is not an algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data inputs [7]. These systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules. For example, an ANN could be trained to identify manufacturing defects by exposing it to example images that have been manually labeled as "defective" or "not defective". Such neural networks are generally trained with little prior knowledge about materials or manufacturing defects. Instead, neural networks automatically generate identifying characteristics from the learning material that they process. In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some nonlinear function of the sum of its inputs. The connections between artificial neurons are called "edges". Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs.

Convolutional Neural Networks: The development of CNNs has led to vast improvements in many image processing tasks. In a CNN, pixels from each image are converted to a featurized representation through series of mathematical operations. Images can be represented as an order 3 tensor $\mathbf{I} \in \mathbb{R}^{H \times W \times D}$ with height H, width W, and D color channels [8]. The input sequentially goes through a number of processing steps, commonly referred to as layers. Each layer i can be viewed as an arbitrary transformation $x_{i+1} = f(x_i; \theta_i)$ with inputs x_i, outputs x_{i+1}, and parameters θ_i. By combining multiple layers it is possible to develop a complex nonlinear function which can map high-dimensional data (such as images) to useful outputs (such as classification labels) [39]. More formally, a CNN can be thought of as the composition of number of functions:

$$f_x = f_N(...(f_2(f_1(x_1;\theta_1);\theta_2)...);\theta_N), \quad (1)$$

where x_1 is the input to the CNN and f_x is the output. There are several layer types which are common to most modern CNNs, including convolution layers, pooling layers and batch normalization layers. A convolution layer is a function $f_i(x_i; \theta_i)$ that convolves one or more parameterized kernels with the input tensor, x_i. Suppose the input x_i is an order 3 tensor with size $H_i \times W_i \times D_i$. A convolution kernel is also an order 3 tensor with size $H \times W \times D_i$. The kernel is convolved with the input by taking the dot product of the kernel with the input at each spatial location in the input. By convolving certain types of kernels with the input image, it is possible to obtain meaningful outputs, such as the image gradients. In most modern CNN architectures, the first few convolutional layers extract features like edges and textures. Convolutional layers deeper in the network can extract features that span a greater spatial area of the image, such as object shapes.

Dataflow Graphs: Many modern machine learning software frameworks represent neural networks as a dataflow graph. In a dataflow graph, the nodes represent units of computation, and the edges represent the data consumed or produced by a computation. Representing a CNN in this form allows the underlying software framework to optimize the execution of math operations through increased parallelism, distributed execution, and compiler-generated optimizations. One way of creating a persistent representation of a neural network is to save the dataflow graph in a standardized machine-readable form. Some of the nodes in the neural network dataflow graph may have parameters θ_i that are optimized when training the CNN. These parameters are generally referred to collectively as model weights. It follows that both the dataflow graph and the associated weights are required to fully represent a trained neural network. An example of a dataflow graph for a single-layer neural network is shown in Figure 1.

In this paper, we propose a standardized representation of CNN based on the dataflow graph model and the existing PMML standard. In our proposed format, each layer is represented as an interconnected node in the dataflow graph. Model weights are

Ferguson, Max; Jeong, Seongwoon; Law, Kincho; Narayanan, Anantha Narayanan; Levitan, Svetlana; Tridivesh, Jena; Lee, Yung-Tsun.
"A Standardized Representation of Convolutional Neural Networks for Reliable Deployment of Machine Learning Models in the Manufacturing Industry."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

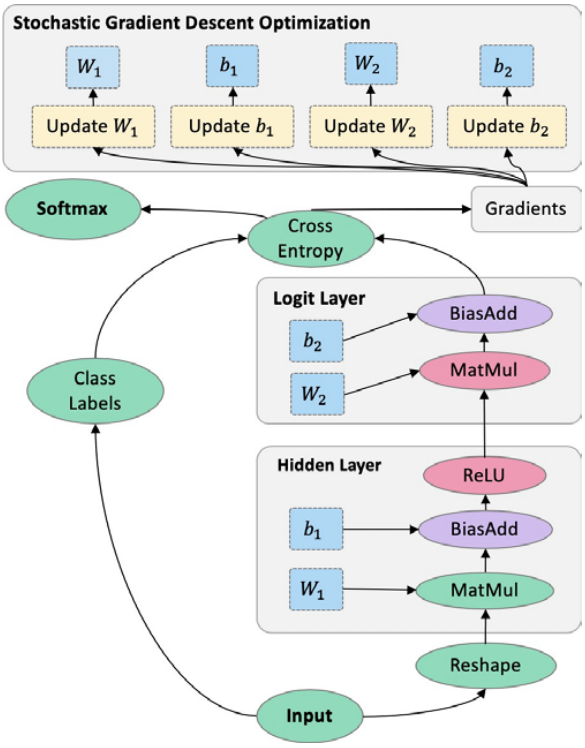


FIGURE 1. DATAFLOW GRAPH FOR TRAINING A SIMPLE NEURAL NETWORK. VARIABLES W_1 AND W_2 ARE THE NEURAL NETWORK WEIGHTS. VARIABLES b_1 AND b_2 ARE THE BIASES.

stored in a separate file, and associated with each node through a unique name. The primary contribution of the paper is the proposed PMML representation for CNN. Additional contributions include a study of the performance characteristics of this proposed format in the context of manufacturing defect detection, and PMML scoring engine.

The remainder of the paper is organized as follows. We begin by exploring related work in the field of standardized models. We then describe our proposed representation, and discuss in detail the mathematical background of each proposed layer. Next, we leverage transfer learning to train ten different CNN models for the task of casting defect detection and represent each trained model in the proposed PMML format. Finally, we develop a high-performance PMML scoring engine for evaluating new input images against PMML models. We show how the scoring engine can be used to evaluate manufacturing images on different hardware devices. The paper is concluded with a brief discussion and conclusion.

RELATED WORKS

There is a large body of work that discusses the use of CNNs for quality control [1,4,8], automated manufacturing [9], and additive manufacturing processes [10]. However, the core focus of this paper is to explore standardized representations for modern CNNs, and demonstrate how such representations can be beneficial to the manufacturing industry. For the remainder of this section, we describe a number of ways that machine learning models are currently stored and distributed.

PMML: PMML provides an open standard for representing data-mining and predictive models [11]. Once a machine-learning model has been trained in an environment like MATLAB, Python, or R, it can be saved as a PMML file. The PMML file can then be moved to a production environment, such as an embedded system or a cloud server. The code in the production environment can parse the PMML file, and use it to generate predictions for new unseen data points. It is important to note that PMML does not control the way the model is trained. PMML is purely a standardized way to represent the trained model.

ONNX: The Open Neural Network Exchange (ONNX) format is a community project created by Facebook and Microsoft [12]. ONNX provides a definition of an extensible dataflow graph model, as well as definitions of built-in operators and standard data types. Each dataflow graph is structured as a list of nodes that form an acyclic graph. Nodes have one or more inputs and one or more outputs. Each node is a call to an operator. Operators are implemented externally to the graph, but the set of built-in operators are portable across frameworks. Every framework supporting ONNX will provide implementations of these operators on the applicable data types.

Keras: Keras is a high-level open source neural network framework written in Python. It is capable of running on top of TensorFlow or Microsoft Cognitive Toolkit. The Keras framework has support for saving and restoring dataflow graphs and model weights. The dataflow graph for any Keras model can be exported to the Javascript object notation (JSON) format. Model weights can be saved in the compact Hierarchical Data format (HDF5) binary format. While these two options provide an easy way to save and distribute Keras models, they do not facilitate interoperability between different tools.

TensorFlow: TensorFlow is an open source software library for high-performance numerical computation. Its flexible architecture allows easy deployment of computation across a variety of platforms. Like Keras, TensorFlow also provides an option for saving dataflow graphs and model weights. The dataflow graph can be exported to a binary file using the Protocol Buffer format. The model weights can be exported to a binary format using a binary format, likely based on the Protocol Buffer format.

PyTorch: PyTorch is an open source machine learning library for Python, based on Torch, used for applications such as natural language processing. It is primarily developed by Facebook's artificial-intelligence research group. PyTorch models are

Ferguson, Max; Jeong, Seongwoon; Law, Kincho; Narayanan, Anantha Narayanan; Levitan, Svetlana; Tridivesh, Jena; Lee, Yung-Tsun.
"A Standardized Representation of Convolutional Neural Networks for Reliable Deployment of Machine Learning Models in the Manufacturing Industry."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

exported to a binary format using the Python Pickle protocol. While this approach provides a seamless method of persisting PyTorch models, it greatly hinders interoperability. In particular, relying on the Pickle object serialization format makes it difficult to transfer PyTorch models to other computing environments that do not support the Python runtime.

Given the current industry practices in machine learning, it appears that there is a need for greater standardization in the representation of deep neural networks. The creation and adoption of such standardized approaches is necessary to promote interoperability of machine learning frameworks and related tools.

PMML FOR CONVOLUTIONAL NEURAL NETWORKS

In this section, we describe how the existing PMML standard is extended to represent CNNs. Figure 2 shows the general structure of a PMML document, which includes four basic elements, namely, header, data dictionary, data transformation, and the data-mining or predictive model [5]. The *Header* element provides a general description of the PMML document, including name, version, timestamp, copyright, and other relevant information for the model-development environment. The *DataDictionary* element contains one or more *DataField* child elements which describe the data fields and their types, as well as the admissible values for the input data. We propose that *DataField* elements with a *dataType* attribute set to "image" can be used to represent images, however further work is needed to formalize this change. In addition, in a classification model, all of the class names are stored in the *DataDictionary* element.

Data transformation is performed using the optional *TransformationDictionary* or *LocalTransformations* element. These elements describe the mapping of the data, if necessary, into a form usable by the mining or predictive model. While support for image transformations would be highly beneficial to the proposed standard, it is outside the scope of this paper, and is left as a topic for future work. The last element in the general structure contains the definition and description of the predictive model. The element is chosen among a list of models defined in PMML standard. We propose the *DeepNetwork* model element as a new element for representing deep neural network models in PMML.

DeepNetwork Element: A CNN model is represented by a *DeepNetwork* element defined in the XML schema, which contains all the necessary information to fully characterize the model. The *DeepNetwork* element has a number of optional attributes that can be used to provide additional metadata about the model, such as the optimization algorithm used to optimize the hyperparameters. Figure 2 shows the elements which can be nested within the *DeepNetwork* element. The *DeepNetwork* element must contain one or more *NetworkLayer* elements which describe individual nodes in the dataflow graph. We now describe a set of layers which are required to minimally represent most of the common CNN architectures that have been proposed to date.

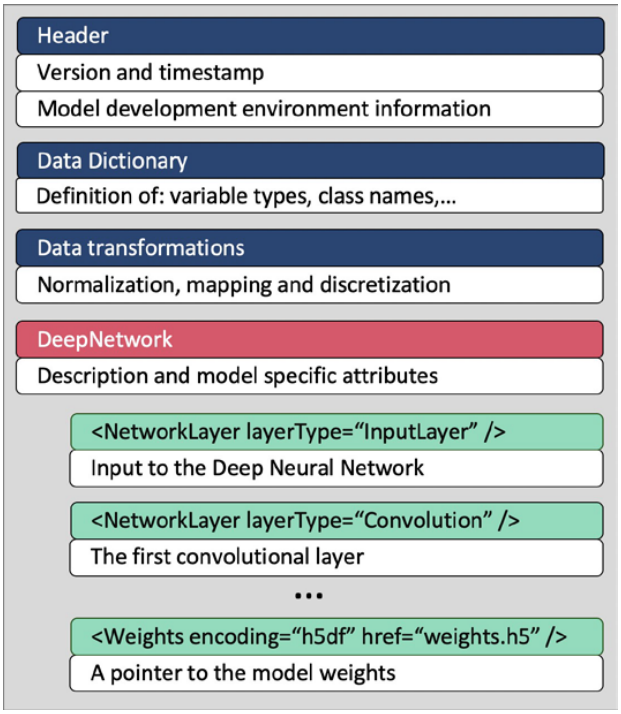


FIGURE 2. THE STRUCTURE AND CONTENTS OF A DEEP-NETWORK PMML FILE.

Convolution Layer: The convolution layer convolves a convolutional kernel with the input tensor. The cardinality of the convolution tensor must be equal to that of the input tensor. The size of the convolutional kernel is governed by the *KernelSize* child element, and the stride is governed by the *KernelStride* child element. An activation function can be optionally applied to the output of this layer. An example of a convolution layer in the proposed PMML format is shown in Figure 3.

Merge Layer: The merge layer takes two tensors of equal dimensions and combines them using an elementwise operator. Allowable operator types are "addition", "subtraction", "multiplication", "division". An example of a merge layer in the proposed PMML format is shown in Figure 4.

Dense Layer: The dense layer represents a fully connected neural network layer. An activation function can optionally be applied to the output of this layer. An example of a dense layer in the proposed PMML format is shown in Figure 5.

Concatenation Layer: The concatenation layer takes two tensors and concatenates them along a given dimension. The cardinality of the two tensors must be equal. The size of all dimensions other than the concatenation dimension must be equal.

Ferguson, Max; Jeong, Seongwoon; Law, Kincho; Narayanan, Anantha Narayanan; Levitan, Svetlana; Tridivesh, Jena; Lee, Yung-Tsun.
"A Standardized Representation of Convolutional Neural Networks for Reliable Deployment of Machine Learning Models in the Manufacturing Industry."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

```
<NetworkLayer
activation="relu"
layerType="Conv2D"
name="block1_conv1"
padding="same"
use_bias="True">
     <InboundNodes>
         <Array n="1" type="string">input_1</Array>
     </InboundNodes>
     <ConvolutionalKernel channels="64">
        <DilationRate>
          <Array n="2" type="int">1 1</Array>
        </DilationRate>
        <KernelSize>
          <Array n="2" type="int">3 3</Array>
        </KernelSize>
        <KernelStride>
          <Array n="2" type="int">1 1</Array>
        </KernelStride>
     </ConvolutionalKernel>
</NetworkLayer>
```

FIGURE 3. CONVOLUTION LAYER EXAMPLE.

```
<NetworkLayer
layerType="Merge"
axis="3"
operator="add"
name="conv2_block1_concat">
  <InboundNodes>
    <Array n="2" type="string">pool1 conv2</Array>
  </InboundNodes>
</NetworkLayer>
```

FIGURE 4. MERGE LAYER EXAMPLE.

```
<NetworkLayer
activation="softmax"
channels="1000"
layerType="Dense"
name="fc1000">
  <InboundNodes>
    <Array n="2" type="string">pool1 conv2</Array>
  </InboundNodes>
</NetworkLayer>
```

FIGURE 5. DENSE LAYER EXAMPLE.

An example of a concatenation layer in the proposed PMML format is shown in Figure 6.

Pooling Layer: The pooling layer applies a pooling operation over a single tensor. The width of the pooling kernel is governed by the *PoolSize* child element, and the stride is governed by the *Strides* child element. The pooling operation can be either "max" or "average" depending on the value of the *operation* attribute.

```
<NetworkLayer
layerType="Concatenate"
axis="3"
name="conv2_block1_concat">
  <InboundNodes>
    <Array n="2" type="string">pool1 conv2</Array>
  </InboundNodes>
</NetworkLayer>
```

FIGURE 6. CONCATENATION LAYER EXAMPLE.

Depthwise Convolution Layer: The depthwise convolution layer convolves a convolutional filter with the input, keeping each channel separate. In the regular convolution layer, convolution is performed over multiple input channels. The depth of the filter is equal to the number or input channels, allowing values across multiple channels to be combined to form the output. Depthwise convolutions keep each channel separate - hence the name depthwise.

Batch Normalization Layer: The batch normalization layer applies a batch normalization operation to the input tensor, which aims to generate an output tensor with a zero mean and unit variance. The linear transformation between input and output is based on the distribution of inputs at test time, and the parameters are generally fixed after training is completed.

Activation Layer: The activation layer applies an activation function to each element in the input tensor. The activation function can be any one of "relu", "sigmoid", "tanh", "selu", "elu", "softmax". The *threshold* attribute allows the activation function to be offset horizontally. The *max_value* attribute limits the maximum value of each output value.

Global Pooling Layer: This layer applies a pooling operation across all spatial dimensions of the input tensor. The pooling operation can be either "max" or "average", and is specified using the *operation* attribute. This layer returns a tensor that has size (*batch_size*, *channels*).

Zero Padding Layer: This layer pads the outside of a 2D tensor with zeros. This operation is commonly used to increase the size of oddly shaped layers, to allow dimension reduction in subsequent layers. The number of zeros that are added to each dimension is specified using the *padding* attribute.

Reshape Layer: The reshape layer reshapes the input tensor. The number of values in the input tensor must equal the number of values in the output tensor. The first dimension is not reshaped as this is commonly the batch dimension.

Flatten Layer: The flatten layer flattens the input tensor such that the output size is (*batch_size*, *n*) where *n* is the number of values in the input tensor.

Many of these layers have weights that are optimized dur-

Ferguson, Max; Jeong, Seongwoon; Law, Kincho; Narayanan, Anantha Narayanan; Levitan, Svetlana; Tridivesh, Jena; Lee, Yung-Tsun.
"A Standardized Representation of Convolutional Neural Networks for Reliable Deployment of Machine Learning Models in the Manufacturing Industry."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

ing the training process. These weights are stored in a binary key-value format, such as HDF5, and the corresponding file is referenced using the *Weights* element. The *Weights* element has *encoding*, *checksum* and *href* attributes which specify the encoding, checksum and location of the model weights,respectively.

Examples of full PMML models for many common CNN architectures are made available at https://github.com/maxkferg/python-pmml [13].

PMML FOR DEFECT DETECTION

In this section, we show how the proposed PMML format can be used to facilitate transfer learning in a defect detection task. CNNs have been particularly successful for detecting manufacturing defects using camera or Xray images [1, 2, 14]. Historically, manufacturing researchers and technicians were concerned that deep learning systems would require an infeasible amount of data to train. However, by leveraging transfer learning, it is now possible to train a powerful computer vision system with as little as a thousand training examples. In transfer learning, a neural network model is first trained on a large dataset, such as the ImageNet dataset. The trained model is then fine-tuned on a smaller domain-specific dataset, such as a casting defect dataset. Hence, the successful application of transfer learning is highly dependent on the availability of high-quality neural network models. Without a standardized representation of neural network models, it is challenging for researchers and practitioners to share their pretrained models.

A total of ten pretrained image classification models are converted to the PMML format, allowing them to be easily loaded into machine learning frameworks such as Keras, TensorFlow or PyTorch. Each model is then fine-tuned on the GDXray casting defect dataset. The fine-tuned models are exported to the proposed PMML format and made publicly available. Finally, the PMML models are loaded onto a TensorFlow scoring engine to demonstrate how the model could be used in a manufacturing environment.

GDXray dataset: The GDXRay dataset is a collection of annotated Xray images [15]. The Castings series of this dataset contains 2727 X-ray images mainly from automotive parts, including aluminum wheels and knuckles. The casting defects in each image are labeled with tight fitting bounding-boxes. The size of the images in the dataset ranges from 256$\times$256 pixels to 768$\times$572 pixels. The GDXray dataset has been used by many researchers as a standard benchmark for defect detection including [16], where patches of size 32$\times$32 pixels are cropped from the GDXray Castings series and used to test a number of different classifiers. The best performance is achieved by a simple local binary pattern (LBP) descriptor with a linear support vector machine (SVM) classifier [16]. Several deep learning approaches are also evaluated, obtaining up to 86.4 % patch classification accuracy. We train multiple different CNN models to classify

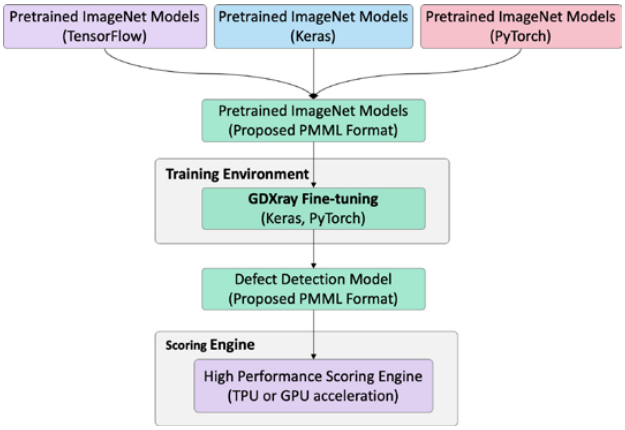


FIGURE 7. TRANSFER LEARNING WITH THE PROPOSED PMML FORMAT. THE PRETRAINED MODELS ARE FIRST CONVERTED TO THE PROPOSED PMML FORMAT AND FINE-TUNED ON THE GDXRAY DEFECT TILE DATASET. THE MODELS ARE THEN TRANSFERRED TO A SCORING ENGINE USING THE PROPOSED FORMAT.

image tiles from the GDXray casting dataset. Each image in the casting set is divided into 224$\times$224 pixel tiles. When an image cannot be divided perfectly into 224$\times$224 pixel tiles, it is first padded with black pixels and then divided into tiles. Image tiles containing one or more casting defects are considered defective ($y = 1$), otherwise the tile is considered satisfactory and assigned class $y = 0$. The goal is to develop a classifier that can correctly predict the class of an unseen tile. To ensure the results are consistent with previous work, the training and testing data is divided in the same way as described in [2]. The preprocessed training dataset contains 8192 tiles. A further 1000 tiles from the training set are assigned to the dev set, which is used to test whether the model is overfitting. The test set contains a total of 1894 tiles.

Machine learning models: Ten different CNN models are trained on the GDXray tiles. Four different CNN architectures are trained, namely the VGG (Visual Geometry Group), Residual Network (ResNet), MobileNet and DenseNet architectures. Pretrained ImageNet models are converted to PMML from the native Keras and PyTorch model formats. The framework-agnostic nature of PMML provides the flexibility to train and save models in any supported language and machine learning framework. The multi-step training process is illustrated in Figure 7. The CNN models are trained on the GDXray dataset using the Keras machine learning framework. Additionally, the VGG and ResNet models are also trained on the GDXray dataset using the PyTorch machine learning framework.

The sparse occurrence of manufacturing defects often creates a challenge when training machine learning models. In

Ferguson, Max; Jeong, Seongwoon; Law, Kincho; Narayanan, Anantha Narayanan; Levitan, Svetlana; Tridivesh, Jena; Lee, Yung-Tsun.
"A Standardized Representation of Convolutional Neural Networks for Reliable Deployment of Machine Learning Models in the Manufacturing Industry."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

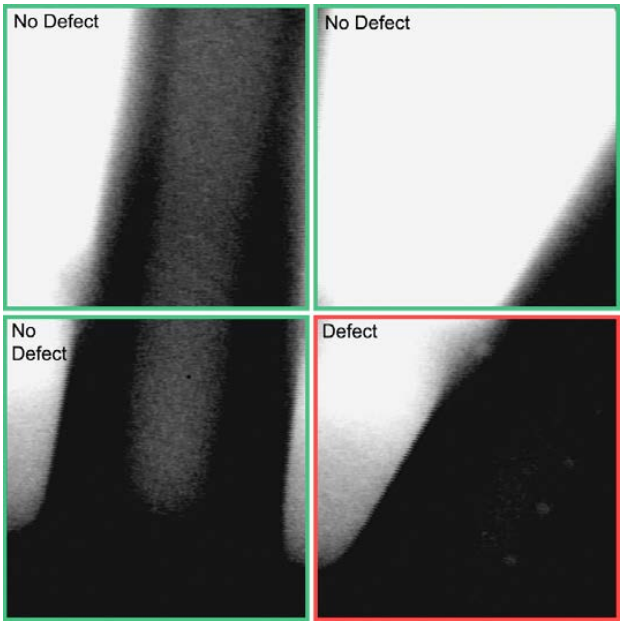


FIGURE 8. PREDICTION RESULTS FOR FOUR TILES FROM THE GDXRAY DEFECT TILE DATASET. EACH 224 × 224 PIXEL TILE WAS CLASSIFIED INDIVIDUALLY, AND THEN THE TILES WERE RECOMBINED TO FORM THE FIGURE.

highly imbalanced datasets, predictive models can often minimize loss by simply predicting the most common class. To avoid this pitfall, defective images are oversampled from the dataset. Specifically, we ensure that at least one defective tile is sampled from the dataset for every three clean tiles. Data augmentation is applied to reduce potential overfitting. During training, each image is rotated 90 degrees with probability 0.5, flipped horizontally with probability 0.5 and flipped vertically with probability 0.5. A small amount of Gaussian random noise is also added to each image. The gradient-based Adam algorithm is used for parameter optimization. The dense layers of each model are trained on the GDXray dataset for 10 epochs with a learning rate of 0.005, while keeping the weights of the convolutional layers fixed. The model is then trained for an additional 10 epochs without fixing any of the weights, using a learning rate of 0.001. After each epoch, the model is saved to PMML and tested on the dev set. The model that achieves the highest performance on the dev set is selected as the final model. The test set prediction accuracy for each model is presented in Table 1. Some example predictions are shown in Figure 8. The trained ImageNet classification models and GDXray defect classification models are made publicly available in the proposed PMML format, to accelerate future research in this direction [13].

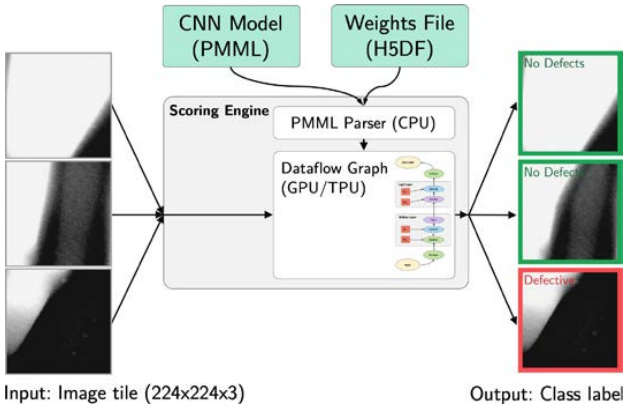


FIGURE 9. NEURAL NETWORK EVALUATION WITH THE PROPOSED PMML FORMAT. A HIGH-PERFORMANCE SCORING ENGINE EVALUATES NEW IMAGES AGAINST THE MODEL AND RETURNS THE PREDICTED CLASS.

EFFICIENCY AND PERFORMANCE

Modern CNNs are growing increasingly complex, with recent architectures having hundreds of layers and millions of parameters. Therefore, storage efficiency, serialization performance and deserialization performance must be considered when designing a standardized format for modern neural networks. To evaluate the performance of the proposed format, we develop a PMML scoring engine with support for deep CNN models, and conduct a number of performance experiments.

There are two main factors when considering the performance of a scoring engine: (1) The amount of time it takes to load a model from a PMML file into memory, and (2) the amount of time required to evaluate a new input. We will refer to (1) as the initial load time and (2) as the evaluation time. The initial load time of modern neural networks tends to be quite large as most modern CNNs have complicated dataflow graphs and millions of parameters. However, models are generally kept in memory between subsequent predictions, so initial load time is not a major influence on performance. The prediction time, however, is critical to many applications in manufacturing, as it dictates the amount of time between an observation being made, and a prediction being obtained.

A scoring engine is developed to evaluate image inputs against CNN models in the proposed PMML format. The scoring engine uses the TensorFlow machine learning framework to perform the mathematical operations associated with each network layer. In our scoring engine, the PMML file is parsed using the lxml XML parser [17] and the Python programming language. The scoring engine is engineered in a way such that the dataflow graph can be evaluated on a central processing unit (CPU), graphics processing unit (GPU), or tensor processing unit

Ferguson, Max; Jeong, Seongwoon; Law, Kincho; Narayanan, Anantha Narayanan; Levitan, Svetlana; Tridivesh, Jena; Lee, Yung-Tsun.
"A Standardized Representation of Convolutional Neural Networks for Reliable Deployment of Machine Learning Models in the Manufacturing Industry."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

TABLE 1. TEST ACCURACY AND MODEL STATISTICS FOR TEN CNN MODELS TRAINED ON THE GDXRAY DEFECT TILE DATASET.

Model	Training Framework	PMML Size (KB)	Weights Size (MB)	Parameters	Training time (s)	Test accuracy
VGG-16	PyTorch	58	528	138,357,544	1442	0.899
VGG-19	PyTorch	60	549	143,667,240	2301	0.921
ResNet-50	PyTorch	106	99	25,636,712	1349	0.933
ResNet-152	PyTorch	154	244	60,344,232	2012	**0.944**
VGG-16	Keras	54	528	138,357,544	1322	0.899
VGG-19	Keras	56	549	143,667,240	1943	0.921
ResNet-50	Keras	101	99	25,636,712	1201	0.933
ResNet-152	Keras	156	244	60,344,232	1894	**0.940**
MobileNet-224	Keras	75	16	4,253,864	1561	0.895
DenseNet-121	Keras	185	33	8,062,504	2104	0.942

(TPU). The scoring engine can be used to evaluate batches of one or more images against a PMML model as shown in Figure 9.

The performance characteristics of the proposed PMML format are analyzed using the aforementioned PMML scoring engine. Experiments are conducted using four common CNN architectures, namely VGG-16, ResNet-152, DenseNet-121 and MobileNet-224. In each case, we use a PMML model trained on the GDXray defect detection task. The experiments are conducted on three platforms: A desktop computer with a single NVIDIA 1080 Ti GPU, an identical desktop computer without a dedicated GPU, and a cloud-based virtual machine with a tensor processing unit (TPU). Figure 10 shows the total amount of time required to build the dataflow graph and load the model weights from file. Figure 11 shows the time required to parse the PMML file and load the dataflow graph into memory, on each of the three platforms. Figure 12 shows the time required to load the neural network weights into memory. Finally, Figure 13 shows the prediction time on the three platforms.

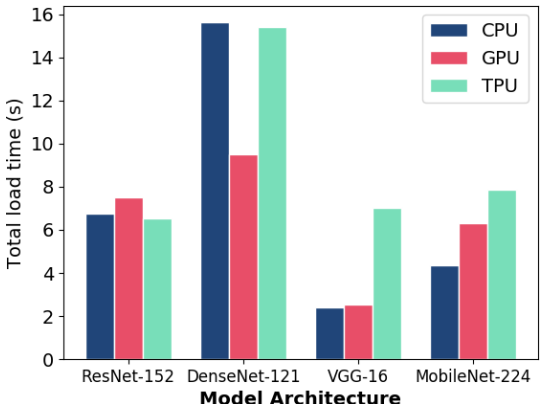


FIGURE 10. TOTAL LOAD TIME FOR FOUR DIFFERENT MODELS. TOTAL LOAD TIME IS DEFINED AS THE TIME FROM WHEN THE PMML FILE FIRST STARTS TO LOAD UNTIL THE TIME THAT THE DATAFLOW GRAPH IS READY TO MAKE PREDICTIONS.

DISCUSSION

In this work, we proposed an extension to the PMML format for the standardized representation of CNN models. The proposed extension adds a DeepNetwork element to the existing standard. The proposed format describes the neural network architecture using the human-readable XML format, making it practical to inspect the architecture of trained models. The human-readable nature of this format is highly beneficial when sharing models trained by researchers, competition teams or large industry teams; rather than documenting the model details in a white-paper, the PMML model can be used to express the architecture in a human-readable manner.

An alternative approach for storing CNN models is to store the entire dataflow graph in a binary format. This approach is currently being employed as part of the ONNX standard. The main benefit of storing the entire dataflow graph is that it pro-

Ferguson, Max; Jeong, Seongwoon; Law, Kincho; Narayanan, Anantha Narayanan; Levitan, Svetlana; Tridivesh, Jena; Lee, Yung-Tsun.
"A Standardized Representation of Convolutional Neural Networks for Reliable Deployment of Machine Learning Models in the Manufacturing Industry."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

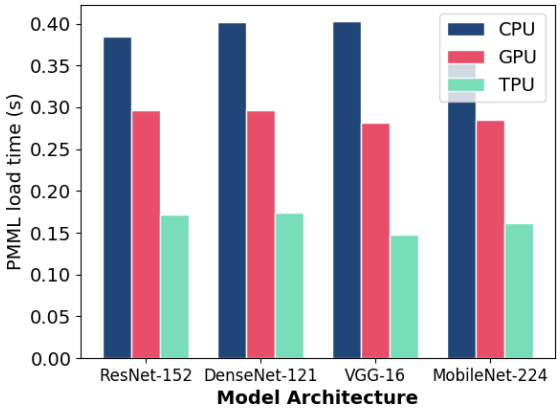


FIGURE 11. PMML LOAD TIME FOR FOUR DIFFERENT MODELS. PMML LOAD TIME IS DEFINED AS THE TIME TAKEN TO LOAD AND PARSE THE PMML FILE.

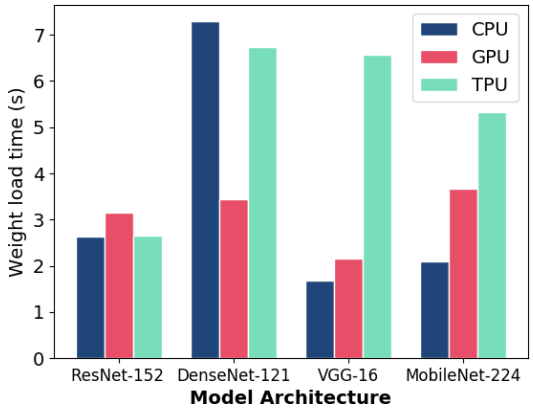


FIGURE 12. WEIGHT LOAD TIME FOR FOUR DIFFERENT MODELS. WEIGHT LOAD TIME IS DEFINED AS THE TIME TAKEN TO BUILD THE DATAFLOW GRAPH ON THE EXECUTION DEVICE AND LOAD THE MODEL WEIGHTS.

vides more flexibility for custom mathematical operations. However, binary formats such as ONNX are not human-readable, making them more difficult to interpret. Both methods could be useful in different contexts: The proposed PMML representation could be particularly useful when widely-used neural network architectures, such as ResNet, are trained on a novel task and shared across the research or industry communities. A binary format such as ONNX is likely more useful for representing complex neural network architectures such as recurrent neural networks.

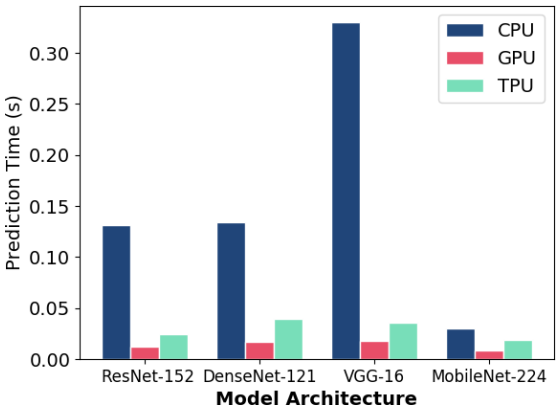


FIGURE 13. MODEL PREDICTION TIME ON THREE DIFFERENT COMPUTATIONAL PLATFORMS. MODEL PREDICTION TIME IS DEFINED AS THE AMOUNT OF TIME REQUIRED TO GENERATE A PREDICTION FROM A NEW INPUT, WHEN USING A BATCH SIZE OF 1.

The neural network weights are critical to fully representing a trained neural network; without these weights a prediction cannot be calculated. While PMML already contains support for a *NeuralNetwork* model element, this requires every neural network node to be specified. Specifying a 224×224 convolution layer with a 3×3 kernel requires 9 weights using the *DeepNetwork* element compared to $224 \times 224 \times 3 \times 3 = 451,584$ weights with the traditional *NeuralNetwork* element. In this work, we chose to use the HDF5 format to store model weights, as it provides a simple and efficient method of storing a map between layer names and weight tensors. However, model weights could also be saved using a binary format like ONNX, or directly embedded in the PMML file as a base64 string.

A tile-base casting defect classifier was developed to illustrate how the proposed PMML format makes transfer learning more accessible to a large number of machine learning frameworks. The ResNet-152 model obtained 94.4 % prediction accuracy on the test set, outperforming the highest scoring CNN model in [16], ImageXnet, which achieved 86.4 % accuracy. The improved performance is likely due to the use of larger tiles (224×224 pixels) compared to the 32×32 pixel patches used in ImageXnet.

Scoring engines will become increasingly important in the smart manufacturing industry, especially as internet-connected manufacturing machines become more mainstream. For many real-time applications, the response time of the scoring engine must be sufficiently low to avoid manufacturing devices becoming idle whilst waiting for feedback from the scoring engine. Due to the declarative nature of the proposed PMML format, it is pos-

Ferguson, Max; Jeong, Seongwoon; Law, Kincho; Narayanan, Anantha Narayanan; Levitan, Svetlana; Tridivesh, Jena; Lee, Yung-Tsun.
"A Standardized Representation of Convolutional Neural Networks for Reliable Deployment of Machine Learning Models in the Manufacturing Industry."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

sible for our scoring engine to evaluate models on CPU, GPU or TPU hardware. This level of flexibility could be highly beneficial for manufacturing applications, where models must be evaluated on a range of hardware systems.

CONCLUSION

In this work, we proposed an extension to the PMML format for a standardized representation of convolutional neural network models. A tile-base casting defect classifier was developed to illustrate how the proposed PMML format makes transfer learning more practical on many machine learning frameworks. A scoring engine was created for this new standardized schema and used to evaluate the performance characteristics of the proposed format. The scoring engine and the trained models have all been made publicly available to accelerate research in the field.

ACKNOWLEDGMENT

This research is partially supported by the Smart Manufacturing Systems Design and Analysis Program at the National Institute of Standards and Technology (NIST), Grant Number 70NANB18H193 awarded to Stanford University.

DISCLAIMER

Certain commercial systems are identified in this paper. Such identification does not imply recommendation or endorsement by NIST; nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST or any other supporting U.S. government or corporate organizations.

REFERENCES

[1] Ferguson, M., Ak, R., Lee, Y.-T. T., and Law, K. H., 2018. "Detection and segmentation of manufacturing defects with convolutional neural networks and transfer learning". *Smart and Sustainable Manufacturing Systems (SSMS)*, **2**(1), Oct.

[2] Ferguson, M., Ak, R., Lee, Y.-T. T., and Law, K. H., 2017. "Automatic localization of casting defects with convolutional neural networks". *IEEE International Conference on Big Data (Big Data 2017)*, IEEE, pp. 1726–1735.

[3] Napoletano, P., Piccoli, F., and Schettini, R., 2018. "Anomaly detection in nanofibrous materials by CNN-based self-similarity". *Sensors*, **18**(1), p. 209.

[4] Chu, Y., Huang, C., Xie, X., Tan, B., Kamal, S., and Xiong, X., 2018. "Multilayer hybrid deep-learning method for waste classification and recycling". *Computational Intelligence and Neuroscience*, **2018**.

[5] Guazzelli, A., Zeller, M., Lin, W.-C., Williams, G., et al., 2009. "PMML: An open standard for sharing models". *The R Journal*, **1**(1), pp. 60–65.

[6] Gorea, D., 2008. "Dynamically integrating knowledge in applications. An online scoring engine architecture". *Advances in Electrical and Computer Engineering*, **8**(15), pp. 44–49.

[7] Hecht-Nielsen, R., 1992. "Theory of the backpropagation neural network". *Neural Networks for Perception*. Elsevier, pp. 65–93.

[8] Hanzaei, S. H., Afshar, A., and Barazandeh, F., 2017. "Automatic detection and classification of the ceramic tiles surface defects". *Pattern Recognition*, **66**, pp. 174 – 189.

[9] Allard, U. C., Nougarou, F., Fall, C. L., Giguère, P., Gosselin, C., Laviolette, F., and Gosselin, B., 2016. "A convolutional neural network for robotic arm guidance using semg based frequency-features". *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp. 2464–2470.

[10] Shevchik, S. A., Kenel, C., Leinenbach, C., and Wasmer, K., 2018. "Acoustic emission for in situ quality monitoring in additive manufacturing using spectral convolutional neural networks". *Additive Manufacturing*, **21**, pp. 598–604.

[11] Zeller, M., Grossman, R., Lingenfelder, C., Berthold, M. R., Marcade, E., Pechter, R., Hoskins, M., Thompson, W., and Holada, R., 2009. "Open Standards and Cloud Computing: KDD-2009 Panel Report". *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 11–18.

[12] Candela, J. Q., 2017. Facebook and Microsoft introduce new open ecosystem for interchangeable AI frameworks. Available at https://fb.me/candela_2017. Accessed May 22, 2019.

[13] Ferguson, M., 2019. Python PMML scoring engine and CNN models. Available at https://github.com/maxkferg/python-pmml/. Accessed February 19, 2019.

[14] Ren, R., Hung, T., and Tan, K. C., 2018. "A generic deep-learning-based approach for automated surface inspection". *IEEE Transactions on Cybernetics*, **48**(3), pp. 929–940.

[15] Mery, D., Riffo, V., Zscherpel, U., Mondragn, G., Lillo, I., Zuccar, I., Lobel, H., and Carrasco, M., 2015. "GDXray: The database of Xray images for nondestructive testing". *Journal of Nondestructive Evaluation*, **34**(4), Nov., p. 42.

[16] Mery, D., and Arteta, C., 2017. "Automatic defect recognition in Xray testing using computer vision". *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp. 1026–1035.

[17] Behnel, S., Faassen, M., and Bicking, I., 2005. lxml: XML and HTML with Python.

Ferguson, Max; Jeong, Seongwoon; Law, Kincho; Narayanan, Anantha Narayanan; Levitan, Svetlana; Tridivesh, Jena; Lee, Yung-Tsun.
"A Standardized Representation of Convolutional Neural Networks for Reliable Deployment of Machine Learning Models in the Manufacturing Industry."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

Advanced Sensor and Target Development to Support Robot Accuracy Degradation Assessment

Guixiu Qiao, *Senior Member, IEEE*

***Abstract*— This paper presents a vision-based, 6 degree of freedom (DOF) measurement system that can measure robot dynamic motions in real-time. A motorized target is designed as a part of the system to work with a vision-based measurement instrument, providing unique features to stand out from the background and enable the achievement of high accuracy measurement. With the capability to measure a robot's 6 DOF information, the robot's accuracy degradation can be monitored, assessed, and predicted to avoid a costly, unexpected shutdown, or decrease in manufacturing quality and production efficiency. The National Institute of Standards and Technology (NIST) is developing the necessary measurement science to support the monitoring, diagnostics, and prognostics of robot systems by providing intelligence to enhance maintenance and control strategies. The robot accuracy degradation research includes the development of modeling and algorithm for the test method, advanced sensor and target development to accurately measure robot 6 DOF information, and algorithms to analyze the data. This paper focuses on the development of the advanced sensor and target. A use case shows the use of the measurement system on a Universal Robot to support the robot accuracy degradation assessment.**

I. INTRODUCTION

Robots are known for their repeatability, but more accurate robots have become valuable tools to enable broader robot applications [1-4]. For example, more off-line programming can be performed because robots can move to desired positions precisely with improved absolute accuracy [2]. Also, many new robot applications such as robot material removal, high precision assembly, robotic drilling, robot riveting, and robot metrology require robots with high accuracy [5-7]. Compared to expensive solutions which use custom machines, high accuracy robots with articulated arms can extend arm to cover a relatively large work volume and can navigate along curvature surfaces or into tight spaces. Implementing robots in manufacturing processes benefits manufacturers by improving flexibility and reducing costs.

As more robotic technologies are integrated into complex manufacturing environments, it is critical to understand a robot system's reliability [10]. A manufacturing system's efficiency, quality, and productivity compromise can be comprised by the robot's accuracy degradation. Robot accuracy degradation is relatively difficult to be detected compared to the hard stop of a production line [8, 9]. Although the robot's performance is degraded, the robot is still running, and parts are making. However, the robot is working at a decreased level of performance. Moreover, a robot's accuracy degradation may also influence other automation components' performance. For example, stresses and strains may accumulate when a fixture or a gripper is working constantly in a biased position. The accumulated stresses and strains may result in the mechanical failure or wear of the fixture or gripper. It is very challenging to decouple errors and find the root causes of failures caused by robot accuracy degradation.

To assess robot accuracy degradation, the deviations of the robot's motion need to be measured. Fig. 1 shows a robot's actual motion has deviated from the designed motion. Moreover, the deviations are different at different locations within the robot's workspace. To determine the deviation, a sensor is needed to measure the x, y, z, pitch, yaw, and roll of the robot tool center point (TCP). As shown in Fig. 1, TCP is located at the end of a robot's kinematic chain. Any error in the robot kinematic chain will be reflected as a deviation of the TCP position and orientation. There is a wide range of techniques being used to measure and calibrate robots. These techniques include: 1) Pose matching methods via driving a robot to a known location, and the pose calculated by the robot controller being recorded [10]; 2) Polar measurement techniques with laser trackers or total stations [11]; 3) Trilateration with a theodolite, using cable potentiometer systems, or laser interferometers [12]; 4) Tactile techniques using gauges or coordinate measurement machines [13]; 5) Inertial navigation systems and magnetic field systems [14]; 6) Photogrammetry with high-resolution digital systems [15-17].

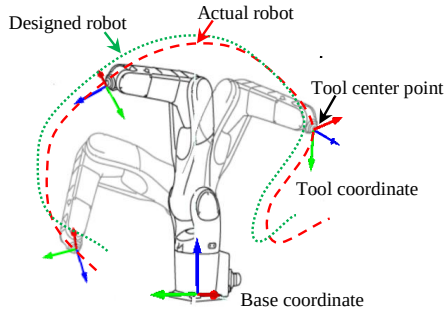


Fig. 1. Robot tool center point

Some of the above-mentioned measurement techniques involve rather expensive metrology instruments (e.g., laser trackers, total stations, etc.). Pose matching, gauges, or coordinate measurement machines are very slow. Trilateration and other methods usually lack orientation information. The vision-based system is gaining more attention in recent years. They have the advantages of being relatively low-cost and non-contact [18]. Innovations in new vision sensors (including low noise, high dynamic range, high resolution, and hardware-accelerated processing) are accelerating the

Guixiu Qiao is with the National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (phone: 301-975-2865; fax: 301-990-9688 e-mail: guixiu.qiao@nist.gov).

Qiao, Guixiu.
"Advanced Sensor and Target Development to Support Robot Accuracy Degradation Assessment."
Paper presented at IEEE International Conference on Automation Science and Engineering (CASE2019), Vancouver, BC, Canada. August 22, 2019 - August 26, 2019.

application of vision-based systems [19]. For these reasons, a pilot study on the applicability of a dual camera stereo system based on low-cost vision hardware components to industrial robot health assessment tasks was conducted. This study focused on aspects of accuracy and real-time processing.

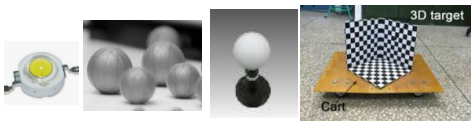
The measurement of the TCP deviation is the first step to assess robot accuracy degradation. A robot's TCP deviation varies at different locations. Even at the same location, if the robot approaches the location from different directions, the TCP deviations are different. Therefore, there is an infinite number of combinations of locations and directions. It is impossible to measure all possible combinations to determine the overall accuracy of the robot. There needs to be a methodology to efficiently measure, monitor, diagnose, predict, and maintain the health of a robot (collectively known as Prognostics and Health Management (PHM)).

The National Institute of Standards and Technology (NIST) is working to develop the measurement science in assessing robot health and optimizing the maintenance of robot systems. As a subset of this research, a quick health assessment methodology was developed to enable manufacturers to quickly assess a robot's accuracy degradation throughout the robot workspace [20]. This paper focuses on the advanced sensor and target development to support the robot's health assessment by examining the degradation of the robot TCP accuracy. The following sections present the design principle of the new target that exceeds the existing target representation; analyze the accuracy and real-time processing capability; show the design of the system; and present a use case of using the system on a Universal Robot to support the accuracy degradation assessment methodology.

II. Design Principle of 3 DOF and 6 DOF Representation

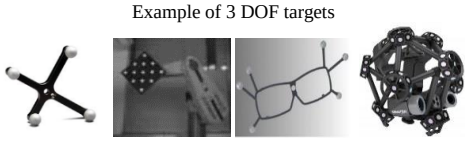
Any possible movement of a rigid body can be expressed as a combination of the basic 6 DOF - 3 translations and 3 rotations. Translation has 3 degrees of freedom: forward/back, up/down, left/right. Rotation has 3 degrees of freedom: pitch, yaw, and roll. A measurement system usually contains the measurement instrument (or sensor) and a measurement target. The measurement target defines what features can be captured by the measurement instrument to represent 3 DOF or 6 DOF information.

The 3 DOF translation is usually represented as a fixed point (x, y, z) on an object. Because once the (x, y, z) is defined, the object is not free to translate in any direction. When the object moves, its translation can be measured via measuring the position changes of the point on the object. If a measurement system can measure a point, the measurement system can measure the 3 DOF information of the object. Fig. 2 (a) to (d) are examples of 3 DOF targets used by vision-based measurement instruments. Fig. 2 (a) is a light-emitting diode (LED) target. It can only be viewed when the LED target is facing the measurement instrument. The output is the center position (x, y, z) of the LED target. Fig. 2 (b) shows sphere targets used by photogrammetric systems. A sphere target is a target that can be measured from any view. The sphere center is the (x, y, z) position to be measured. A photogrammetric system captures the two dimensional (2D) image of the sphere target. The centroid of the 2D image is detected and later triangulated to a 3 DOF point that represents the sphere center. Fig. 2 (c) is an example of the sphere target used by scanning systems. A scanning system scans the surface of the sphere and outputs a point cloud. A sphere surface is constructed using a best-fit method. Then the center of the sphere is calculated. Fig. 2 (d) is the other type of 3 DOF target used by vision-based systems. The intersection corner of the checkerboards is the point being measured.



(a) (b) (c) (d)

Example of 3 DOF targets



(e) (f) (g) (h)

Example of 6 DOF targets

Fig. 2. 3 DOF and 6 DOF target representation

Besides translation, many applications also need rotation information. When the 3 degrees of rotation is added to the translation, a coordinate frame is formed. The origin of the coordinate frame represents the point where translation is measured. Positional tracking can be performed by tracking the origin position. To define rotation, three axes of the coordinate frame are used. They are: a primary axis, a secondary axis, and a tertiary axis. In Fig. 1, the tool frame is made up of the three axes at the TCP. The tool center point is the origin of the frame. The tool coordinate is programmable and can be "taught" for each tool or fixture attached to the robot. Any error in the kinematic chain will be reflected as the TCP error. Measuring the 6 DOF errors of the tool frame is a measure of robot accuracy. If a measurement system can measure a coordinate frame, the measurement system can measure the 6 DOF information of the object.

Existing 6 DOF target representation can be created by combining multiple 3 DOF targets. One representation that has been widely used to define a coordinate frame is using the three point method. A coordinate frame is defined using three points. A coordinate frame is represented by defining an origin and two axes. The third axis is perpendicular to the other two axes. The three points are used as: 1) a point defining the origin, 2) a point defining the primary axis, e.g., X-axis; this axis is formed by creating a vector from the origin pointing to the point, and 3) a point in a plane, e.g., XY plane. The secondary axis is in this XY plane. It is defined by a vector that starts from the origin and is perpendicular to the X-axis. The third point is used to define the secondary axis's positive direction. Examples of 6 DOF targets are shown in Fig. 2 (e) to (h). In Fig. 2 (e) and (f), multiple spheres use a planar layout to construct a coordinate frame. More than three sphere targets are implemented to create redundancy. The need for redundancy is to avoid image overlapping when viewing the target from different viewpoints. In Fig. 2 (g), spheres use a spatial layout to construct a frame. Fig. 2 h) shows a complex

Qiao, Guixiu.
"Advanced Sensor and Target Development to Support Robot Accuracy Degradation Assessment."
Paper presented at IEEE International Conference on Automation Science and Engineering (CASE2019), Vancouver, BC, Canada. August 22, 2019 - August 26, 2019.

LED spherical array. It combined multiple LEDs to create a 6 DOF target.

For robot accuracy degradation assessment task, a measurement system needs to perform accurate measurements and measure the TCP's 6 DOF information while the robot is stationary or moving. A novel measurement system is developed at NIST to achieve high accuracy and high speed 6 DOF measurements. A dual-camera measurement instrument and a smart target construct a measurement system. The following sections of the paper describe its accuracy and real-time process potential by analyzing the feature detection uncertainty and coordinate frame construction uncertainty.

III. New Target Development

NIST developed a 6 DOF measurement system was developed to perform the assessment of the industrial robot accuracy degradation. The system consists of a measurement instrument with two high-speed color cameras and a smart target as shown in Fig. 3. The smart target is mounted at the end of the robot arm. The vision-based measurement instrument is mounted on a tripod and placed on the floor facing the robot. A measurement software was developed to take measurements of robot TCP position and orientation.

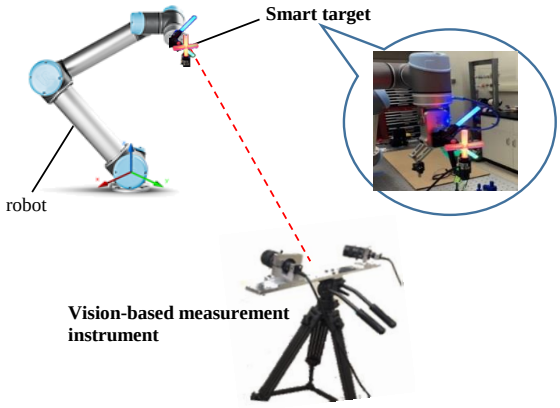
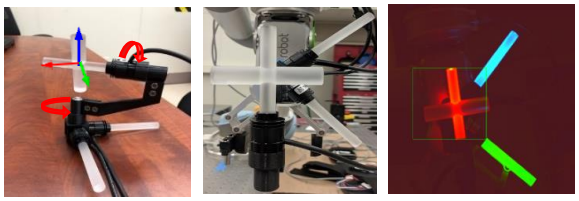


Fig. 3. Sensor and target development to support robot accuracy degradation assessment

The vision-based system is used for the measurement system because: (1) A vision-based system is an non-contact measurement system that can capture both position and orientation simultaneously; (2) High accuracy measurement (with sub-pixel accuracy) has been enabled by novel camera and image processing technology; and (3) A vision-based system is relatively cost-effective to integrate [18] with the mature of camera technology. Color images provide redundant information to get more accurate target detection. The use of high-speed camera and real-time computation enable high-speed measurements. The output of the measurement is (x, y, z, pitch, yaw, and roll) of a moving object with high accuracy.

The smart target is protected under a U.S. provisional patent. It contains light pipes illuminated in three colors. The 6 DOF information is represented by a coordinate frame. As shown in Fig. 4, two cylindrical light pipes form an intersection that represents the origin. The coordinate frame of the smart target is shown in Fig. 4 (a). The two intersecting light pipes are mounted on two motorized rotation axes. The two rotary axes are motorized. Driven by inertial measurement sensors, the intersection pipes can rotate constantly so that the target Y axis points towards the measurement instrument. This creates a constant line-of-sight between the target and the measurement system, even when the target is moving. On the bottom of the smart target, the directions of the x and y axes are defined by two different color light pipes. Fig. 4 (b) shows the smart target mounted on the robot end effector. Fig. 4 (c) shows the light pipe images as seen by the camera when the light pipes are illuminated.



(a) (b) (c)

Fig. 4. Smart target developed at NIST

IV. Analysis on Accuracy and Real-time Processing Capability

To analyze the accuracy of the 6 DOF representation, the sources of measurement uncertainty need to be known. The 3/6 DOF measurement is a mix of hardware and software to detect the position (and orientation) of an object. A vision-based measurement instrument takes images of the target, performs image processing to capture features, and finally outputs the measurement result in either 3 or 6 DOF. Two major concerns of vision-based measurement systems are accuracy and real-time processing capability. Existing targets have different uncertainties in the detection of features when the target is at different distances or at different poses. These uncertainties influence the accuracy of the final measurement. Also, ambient light has strong effects on the image quality and affects the robustness and accuracy. When the target is in an industrial environment with a complex background, the ability to isolate the target from the background influences the efficiency of real-time processing to track a moving object.

A. Uncertainty calculation for feature detection

Features are components used to construct a coordinate frame. A point feature is a basic feature to represent the translation or the origin of a coordinate frame. The uncertainty of the point feature detection influences the final measurement accuracy. A very common target artifact to define a point is a sphere target. The sphere target can be measured from different views and the derived sphere center is a point feature.

The roundness of the sphere target and the evenness of the surface influence the result of the sphere center detection. For example, Fig. 2 (b) shows the reflective spheres used by infrared cameras. The sphere roundness can be well controlled by machining. However, to make a sphere reflect infrared radiation (IR) light, there are two methods to create the target. The more accurate method is to apply the coating with reflective material. It is expensive in manufacturing because the manufacturing process needs high accuracy in coating control. Reflective tape can be wrapped around a sphere to create the reflective sphere target. It is challenging to

Qiao, Guixiu.
"Advanced Sensor and Target Development to Support Robot Accuracy Degradation Assessment."
Paper presented at IEEE International Conference on Automation Science and Engineering (CASE2019), Vancouver, BC, Canada. August 22, 2019 - August 26, 2019.

control the roundness of the wrapped spherical. Thus, this kind of target can be used for moderate accuracy applications.

The calculation of sphere centers also has many uncertainties. For scanning systems, since the output of the sphere surface is point cloud, the best-fit method is used to construct the sphere. The sphere center is then calculated. If the point cloud consists of the whole sphere surface, the sphere center can be accurate. But in real situations, only a partial sphere surface will be captured. The best-fit result is biased in this condition. Moreover, the uncertainty varies when measuring the sphere from a different view. For photogrammetric systems, the sphere target is captured as 2D images which are circulars on the camera sensors. The centroid of the circle needs to be calculated. The centroid detection accuracy is influenced significantly by the image quality, such as the image exposure, image focus, and the number of the image pixels that represent the target in the camera sensor. The ambient light also has a strong influence on the uncertainty of sphere center detection.

The smart target developed at NIST does not use a sphere to define the coordinate origin. Instead, line features are used. Two perpendicular lines form a "cross". The intersection of the lines defines the coordinate frame origin. The line features are created using cylindrical light pipes. The dimension of the light pipe artifact is 10 mm in diameter and 75 mm long. Three colors of LEDs are used in the smart target design with special wavelengths that match with the narrow band filters on the measurement instrument's cameras. The purpose is to reduce the effects of ambient light. With special surface finishing, the entire cylindrical surface has an even light distribution. Fig. 5 shows the image of light pipe artifacts representing line features.

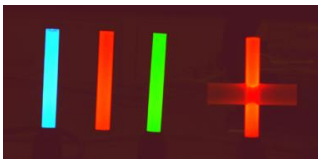


Fig. 5. Light pipe artifacts of smart target for line feature representation

B. *Uncertainty calculation for coordinate frame construction*

After the basic features are detected, they are used to construct a coordinate frame. When using sphere targets to construct a coordinate frame, one of the sphere centers is selected as the origin. The problem of this method is: the origin definition inherits the uncertainty from the sphere center measurement in the scale of one-to-one.

For axis definition, the axis of a coordinate frame is defined using two points (two sphere centers) – the origin and a point on the axis. Fig. 6 shows the axis definition error caused by point detection uncertainty. P0 is the origin point. P1 is the true position of the second point. P1' is the real P1 position with Delta error caused by the uncertainty of point P0 and P1. The angular error of the axis definition is $\tan^{-1}(Delta/\overline{p0p1})$. Even a small positional error of the point (Delta) can cause a large angular error of the axis vector.

For the same Delta error, a shorter distance between P0 to P1 corresponds to a larger angular error. In Fig. 6, P2' has the same Delta error, but the angular error ***a*** is bigger than the angular error ***b*** because P1 is closer to P0 than P2. Therefore, to achieve higher accuracy, the target needs to be larger to maximize the distance $\overline{p0p1}$. However, it is not practical to build a very large target. As a result, for a target that uses two/multiple spheres to define the axis, the measurement has larger angular uncertainties than a target using line features.

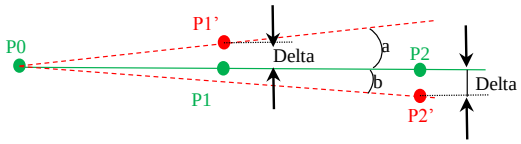


Fig. 6. Axis definition error caused from point detection uncertainty

Besides the direct definition of a coordinate frame, some technologies use the best-fit method to find the transformation between the same points measured in two coordinate frames. Fig. 7 (a) shows a group of points in one coordinate frame and Fig 7. (b) shows the same point group in another coordinate frame. The point group in Fig. 7 (b) is transformed to the coordinate frame in Fig. 7 (a), and the best-fit result is shown in Fig. 7(c). Best-fit usually uses the method of minimizing the sum of the squares of the offsets as the cost function. This method eliminates the one-to-one error propagation from the point detection to the final result. However, the best-fit method only minimizes the translation deviation. Angular errors are not minimized and are, therefore, less accurate.

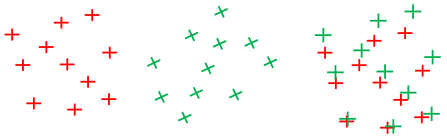


Fig. 7. Uncertainty of sphere center

To avoid these problems, the smart target uses line targets instead of sphere targets. Cylindrical artifacts are used to create line features. The center line of the cylindrical artifact used in the smart target defines an axis. Fig. 8 shows an example of line construction. Since more points are used to fit the axis line, the accuracy of the line detection is much higher compared with the method of defining an axis using two sphere centers. Also, with more points used to fit the line, algorithms can filter out outliers for better line construction. Additionally, the light pipe consists of extra features including color, width, edge features, etc. They can be used to improve the accuracy of target detection. Moreover, best-fit can output multiple results for best-fit data analysis, including straightness, error distribution, error pattern, etc. The best-fit straightness of the line can be used as an important factor to monitor the camera lens distortion. If the straightness of the line fitting results shows distortion, the camera and lens need to be checked. The design of the smart target enables the on-site self-checking capability of the measurement system.

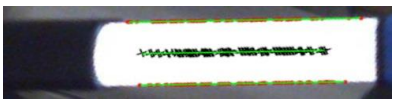


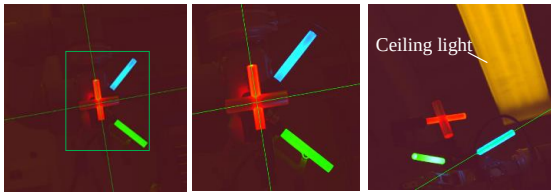
Fig. 8. Center line construction image for a cylinder artifact

Qiao, Guixiu.
"Advanced Sensor and Target Development to Support Robot Accuracy Degradation Assessment."
Paper presented at IEEE International Conference on Automation Science and Engineering (CASE2019), Vancouver, BC, Canada. August 22, 2019 - August 26, 2019.

C. *Method to stand out an object from the background*

The lighting condition in an industrial environment is complex. Various light sources, including ceiling lights, light towers for safety systems, and LED indicators on fixtures, may influence the image quality of the vision-based measurement system. A widely used technology to make a target stand out in a complex background is the utilization of infrared (IR) technology. The shortcoming of this method is the images contain only the markers. When ambient lights exist in the environment, the reflected light from ambient objects will be treated as real targets. There is no redundancy to judge if the detected target is a fake target when applications are used in a complex industrial environment.

To avoid the shortcoming of IR images, high-speed color cameras were selected for the measurement system developed at NIST. Fig. 9 shows the color image of the smart target. The smart target uses three colors of LEDs to illuminate the light pipes. The wavelength combination of the three colors is used as the "signature" of the smart target. Using this "signature", the target can be quickly identified from the complex background. Once the target is found, a bounding box is defined surrounding the target and will be constantly tracked when the target moves with the robot arm. The camera will use a feature of Area of Interest (AOI). It allows a camera using a small AOI window size to operate at higher frame rates compared with one using a full frame capture. Since a smaller window size requires less calculation time compared to a full window size, more complex algorithms can be applied to achieve more robust and accurate results. Fig. 9 (a) shows that the smart target was identified. Fig. 9 (b) shows that the cross center was detected to define the coordinate frame origin. Fig. 9 (c) shows the axis line is detected even when the ceiling light is within the view of the cameras. The color information is used to speed up the calculation and provide more redundancy in the algorithms, for example, avoidance of fake target detection. Advanced color image processing (e.g., pattern reorganization) techniques are utilized to get more accurate target detection results. A graphics processing unit (GPU) technique is used to accelerate the image processing for real-time measurement. The output of the measurement is (x, y, z, pitch, yaw, and roll) of a moving object.



(a) (b) (c)

Fig. 9. Color image of the smart target

In summary, besides the advantages in accuracy improvements, the measurement system has advantages in solving the following challenges required for robot dynamic accuracy assessment. These challenges include:

1) *Dynamic measurement*

Since the smart target measures 6 DOF, every snap of the smart target gives the position and orientation. The high-speed camera can take 175 frames per second in full frame mode. The camera speed can be increased further when using the camera AOI mode to track the small area around the target, allowing the dynamic movement of the robot to be captured.

2) *Non-blocking measurement design*

Traditional targets have an image overlapping problem. The target may block itself in some poses. The smart target is motorized by rotating on two rotary gimbals. the target always rotates towards the measurement system. This eliminates self-blocking and yields optimized pose for measurement.

3) *Unique definition of a coordinate frame*

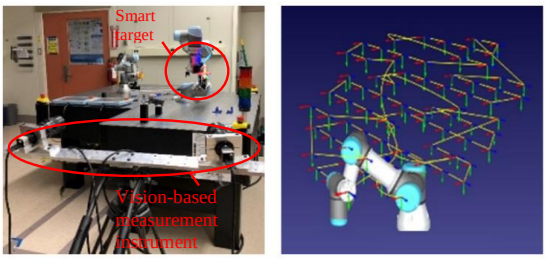
Traditional spherical targets can use any sphere as the origin. It is hard to find the coordination definition when multiple coordinates exist in a system. The 6 DOF smart target uses the cross center as the origin and other two light pipes as the axis direction. This creates a consistent frame definition.

V. Use Case Development

A quick health assessment methodology is developed at NIST. The purpose is to assess the robot accuracy degradation throughout the robot workspace. The quick health assessment methodology includes: 1) development of a sensor and target to measure the robot (x, y, z, pitch, yaw, and roll of the TCP); 2) a robot error model to represent the robot's geometric and non-geometric errors; 3) a test method to define the robot movements; and 4) algorithms to process the measured data to assess the robot's health status. In the quick health assessment methodology, step 2) develops an error model to represent a robot's deflections of the robot's structure and joints, the ideal and non-ideal motion of joints. Step 3) generates a measurement plan that satisfies the requirements to support the robot error model identification. The robot is commanded to move based on the measurement plan. The movements are measured by the advanced sensor and target. Measurement data is fed to the algorithm developed in step 4) to assess the robot health and predict the failure of robot operations under the current accuracy status [9].

The developed measurement system is used to acquire the robot TCP 6 DOF information. As shown in Fig. 10 (a), the smart target is mounted at the end of a Universal Robot arm. The cross light pipe is motorized to constantly rotate toward the camera instrument as shown in Fig. (a). The vision-based measurement instrument is mounted on a tripod and set up at the opposite end of the robot. Fig. 10 (b) shows the measurement plan generated for the Universal Robot. The measurement plan requires the robot TCP to move throughout the entire workspace. The motions distribute evenly in both joint space and Cartesian space. The even distribution of sampling prevents the analysis algorithm from missing errors or too heavily weighting errors, which will bias the results. The coverage of the overall joint space and Cartesian space means that the measurement plan will exercise the robot beyond a partial range of joints or work zones. The coverage of overall joint space enables the capture of joint performance through the full motor and encoder ranges. Covering the entire workspace enables the evaluation of various rigidity conditions. To minimize potential interruptions during robot motion and measurement, a collision check is made during the measurement plan generation process. Also, a line-of-sight check is performed to ensure the planned positions do not occlude the target from the measurement instrument, i.e., arm blocking the target. The output of the measurement system is x, y, z, pitch, yaw, and roll.

Qiao, Guixiu.
"Advanced Sensor and Target Development to Support Robot Accuracy Degradation Assessment."
Paper presented at IEEE International Conference on Automation Science and Engineering (CASE2019), Vancouver, BC, Canada. August 22, 2019 - August 26, 2019.



(a) (b)

Fig. 10. Robot quick health

The TCP 6 DOF measurements are the input of the test method model and the analysis algorithms. The test method model builds a robot error model that can represent a robot's position dependent, non-geometric errors. A novel algorithm is developed to solve the robot error model that contains hundreds of unidentified parameters [9]. A method is developed to decouple the uncertainties of the measurement instrument from the actual robot errors. The analysis outputs the accuracy degradation assessment results.

The quick health assessment can be used to swiftly detect degradations in robot accuracy by finding the robot pose deviations from the nominal poses. The methodology provides manufacturers a tool to quickly detect problems in scenarios when the environmental conditions have changed, reconfigurations are needed, or a critical task is about to perform. The quick health assessment methodology can help to reduce unexpected shutdowns, and help optimize the maintenance strategy to improve productivity via monitoring the robot performance degradation.

VI. Summary

This paper presents the development of an advanced sensor and target to assess robot accuracy degradation. The sensor and target (U.S. patent pending for the smart target) are featured with novel designs that differ from and exceed the performance of existing vision-based measurement systems, especially with respect to accuracy and real-time processing potential. The use of line features enables the high accuracy measurement of the 6 DOF information. The smart target can be utilized in a variety of applications. These applications include registering multiple machines/tools/objects, adaptively locating objects during mobile operations, and precisely tracking the pose of an object. This paper also presents a use case for using the measurement system in assessing robot accuracy degradation. Future efforts are underway to develop additional industrial use cases for applications that require high-precision motions.

NIST Disclaimer

Certain commercial entities, equipment, or materials may be identified in this document in order to illustrate a point or concept. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

References

[1] A. D. Pham and H. J. Ahn, "High precision reducers for industrial robots driving 4th industrial revolution: state of arts, analysis, design, performance evaluation and perspective," *International Journal of Precision Engineering and Manufacturing-Green Technology,* vol. 5, pp. 519-533, Aug 2018.

[2] M. Placzek and L. Piszczek, "Testing of an industrial robot's accuracy and repeatablity in off and online environment," Eksploatacja I Niezawodnosc-Maintenance and Reliability, vol. 20, pp. 455-464, 2018.

[3] H. J. Kim, A. Kawamura, Y. Nishioka, and S. Kawamura, "Mechanical design and control of inflatable robotic arms for high positioning accuracy," Advanced Robotics, vol. 32, pp. 89-104, 2018.

[4] D. Culla, J. Gorrotxategi, M. Rodriguez, J. B. Izard, P. E. Herve, and J. Canada, "Full Production Plant Automation in Industry Using Cable Robotics with High Load Capacities and Position Accuracy," in Robot 2017: Third Iberian Robotics Conference, Vol 2. vol. 694, A. Ollero, A. Sanfeliu, L. Montano, N. Lau, and C. Cardeira, Eds., ed Cham: Springer International Publishing Ag, 2018, pp. 3-14.

[5] D. D. Chen, P. J. Yuan, T. M. Wang, Y. Cai, and L. Xue, "A Compensation Method for Enhancing Aviation Drilling Robot Accuracy Based on Co-Kriging," International Journal of Precision Engineering and Manufacturing, vol. 19, pp. 1133-1142, Aug 2018.

[6] A. Drouot, R. Zhao, L. Irving, D. Sanderson, and S. Ratchev, "Measurement Assisted Assembly for High Accuracy Aerospace Manufacturing," IFAC Papersonline, vol. 51, pp. 393-398, 2018.

[7] A. Klimchik and A. Pashkevich, "Robotic manipulators with double encoders: accuracy improvement based on advanced stiffness modeling and intelligent control," IFAC Papersonline, vol. 51, pp. 740-745, 2018.

[8] G. Liu, "Control of robot manipulators with consideration of actuator performance degradation and failures," in 2001 IEEE International Conference on Robotics and Automation, Vols I-Iv, Proceedings, 2001, pp. 2566-2571.

[9] G. X. Qiao and B. A. Weiss, "Quick health assessment for industrial robot health degradation and the supporting advanced sensing development," Journal of Manufacturing Systems, vol. 48, pp. 51-59, Jul 2018.

[10] Y. Wu, A. Klimchik, S. Caro, B. Furet, and A. Pashkevich, "Geometric calibration of industrial robots using enhanced partial pose measurements and design of experiments," Robotics and Computer-Integrated Manufacturing, vol. 35, pp. 151-168, Oct 2015.

[11] Y. B. HuangFu, L. B. Hang, W. S. Cheng, L. Yu, C. W. Shen, J. Wang, et al., "Research on Robot Calibration Based on Laser Tracker," in Mechanism and Machine Science. vol. 408, X. Zhang, N. Wang, and Y. Huang, Eds., ed Singapore: Springer-Verlag Singapore Pte Ltd, 2017, pp. 1475-1488.

[12] I. A. Sultan and J. G. Wager, "Simplified theodolite calibration for robot metrology," Advanced Robotics, vol. 16, pp. 653-671, 2002.

[13] N. Zaimovic-Uzunovic and S. Lemes, "Cylindricity Measurement on a Coordinate Measuring Machine," in Advances in Manufacturing, A. Hamrol, O. Ciszak, S. Legutko, and M. Jurczyk, Eds., ed Cham: Springer International Publishing Ag, 2018, pp. 825-835.

[14] E. Pivarciova, P. Bozek, Y. Turygin, I. Zajacko, A. Shchenyatsky, S. Vaclav, et al., "Analysis of control and correction options of mobile robot trajectory by an inertial navigation system," International Journal of Advanced Robotic Systems, vol. 15, p. 15, Jan 2018.

[15] B. Diewald, R. Godding, and A. Henrich, Robot calibration with a photogrammetric online system using reseau-scanning-cameras vol. 2252. Bellingham: Spie - Int Soc Optical Engineering, 1994.

[16] J. Peipe, Photogrammetric performance evaluation of digital camera backs for in-studio and in-field use vol. 2598. Bellingham: Spie - Int Soc Optical Engineering, 1995.

[17] A. Filion, A. Joubair, A. S. Tahan, and I. A. Bonev, "Robot calibration using a portable photogrammetry system," Robotics and Computer-Integrated Manufacturing, vol. 49, pp. 77-87, Feb 2018.

[18] M. Švaco, B. Šekoranja, F. Šuligoj, and B. Jerbić, "Calibration of an Industrial Robot Using a Stereo Vision System," Procedia Engineering, vol. 69, pp. 459-463, 2014.

[19] R. Ahmad and P. Plapper, "Safe and Automated Assembly Process using Vision Assisted Robot Manipulator," Procedia CIRP, vol. 41, pp. 771-776, 2016.

[20] G. Qiao, C. Schlenoff, and B. A. Weiss, "Quick positional health assessment for industrial robot prognostics and health management (PHM)," in 2017 IEEE Inter*national Conference on Robotics and Automation (ICRA)*, 2017, pp. 1815-1820.

Qiao, Guixiu.
"Advanced Sensor and Target Development to Support Robot Accuracy Degradation Assessment."
Paper presented at IEEE International Conference on Automation Science and Engineering (CASE2019), Vancouver, BC, Canada. August 22, 2019 - August 26, 2019.

The 27th International Input-Output Association Conference

Quantifying Macroeconomic Resilience Dividends in Cedar Rapids

Juan Fung,† Jennifer Helgeson, Cheyney O'Fallon, David Webb
National Institute of Standards and Technology (NIST)

Harvey Cutler
Colorado State University

Abstract: Cedar Rapids, Iowa offers a unique case study in planning for increased resilience. In 2008, Cedar Rapids experienced severe flooding. Rather than simply rebuilding, the city of Cedar Rapids began to invest in a resilient flood control system and in the revitalization of its downtown neighborhood. This paper develops a Computable General Equilibrium (CGE) model for the regional economy of Cedar Rapids to quantify the net co-benefits of investing in increased resilience, or the "resilience dividend." The resilience dividend includes benefits to the community *even if another disaster does not occur*. The CGE approach to quantifying the resilience dividend can capture how co-benefits are distributed throughout the economy. Our CGE model combines a broad range of data sets, including firm-level employment and wages and property tax assessments, as well as the US Census' Public Use Microdata Sample (PUMS) and US Input-Output tables. We build a CGE model of Cedar Rapids at two different time periods: one in 2007, before the flooding, and one in 2015, after the flooding and initial investment in resilience. We show that a positive productivity shock to the economy results in larger co-benefits for employment and output growth in 2015 than in 2007. The two models demonstrate how economies that invest in increased resilience respond relative to those that do not.

† Corresponding author: juan.fung@nist.gov.

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

1. Introduction

Cedar Rapids, Iowa offers a unique case study in planning for increased resilience. In 2008, Cedar Rapids experienced severe flooding. Rather than simply rebuilding, the city of Cedar Rapids began to invest in both flood mitigation and in the revitalization of its downtown neighborhood. In addition to investments into a resilient flood-control system—a 20-year project that includes levees, removable walls, and new pump stations[1]—the city of Cedar Rapids has also invested in the revitalization of the downtown area in order to have a more dynamic local economy that can absorb shocks, such as extreme flooding, more easily. In addition to making Cedar Rapids more resilient to natural disasters, revitalization of the downtown area also provides benefits for the local economy and social systems in the absence of a natural disaster.

A natural question is whether these investments can produce "co-benefits," such as property value appreciation and business growth. Fung and Helgeson (2017) define the *resilience dividend* as the net co-benefits from investing in increased resilience in the absence of a natural disaster. The resilience dividend encompasses a broader set of potential benefits that can alter how decision makers view return on investment. A positive resilience dividend can help decision makers make a "business case" for resilience. Accounting for co-benefits of resilience planning allows long-term investments to be weighed against day-to-day benefits to the local economy.

This paper presents a Computable General Equilibrium (CGE) model approach to quantifying the resilience dividend. The CGE approach to quantifying the resilience dividend can show at a high level how co-benefits are distributed throughout an economy. The CGE approach is applied to modeling the regional economy of Cedar Rapids. To quantify the resilience dividend from investing in increased resilience, we build a CGE model of Cedar Rapids at two different time periods: one in 2007, before the flooding, and one in 2015, after the flooding and initial investment in resilience. After simulating an increase in productivity in each time period, we find that the increase in employment and output is an order of magnitude larger in 2015 than in 2007. The additional co-benefits in 2015, which are obtained *in the absence of a disaster*, comprise the resilience dividend.

In Section 2, we provide background on the resilience dividend and on our case study, Cedar Rapids. Section 3 provides an overview of the CGE approach to quantifying the resilience dividend, while Section 4 discusses the data we use to build our CGE model. Section 5 presents the CGE model of Cedar Rapids and Section 6 presents the simulation results. Finally, we discuss future directions in Section 7.

2. Background: The resilience dividend in Cedar Rapids

2.1. Background on the resilience dividend

The concept of the resilience dividend was popularized in Rodin (2014), which presents qualitative examples from the real world to illustrate the concept. A series of World Bank reports presented the resilience dividend as arising from a "Triple Dividend of Resilience" largely relevant to disaster risk management (DRM); see Tanner et al. (2016) and Mechler et al. (2016).[2] Bond et al. (2017) describe a

[1] See http://www.cedar-rapids.org/local_government/departments_g_-_v/public_works/cedar_river_flood_control_system.php for an overview and progress report.
[2] The triple dividend consists of: 1. avoided or reduced losses, in the event of a disruptive event occurring; 2. increased economic resilience from reduced disaster risk; and 3. co-benefits for development.

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

Resilience Dividend Valuation Model (RDVM) and present six case studies in the developing-country context to illustrate.[3]

Fung and Helgeson (2017) note that much of the research to date on co-benefits focuses on climate change mitigation and adaptation. Moreover, co-benefits of resilience planning are typically considered in a developing-country context. Finally, quantification of co-benefits is very limited. This is understandable, as it is difficult to determine the full range of co-benefits *ex ante*, as in a Benefit-Cost Analysis (BCA), and to fully track co-benefits flowing throughout an economy *ex post*, as in a CGE model. Nevertheless, the CGE approach can provide a broad picture of how co-benefits are *distributed* throughout an economy.

2.2. The 2008 floods in Cedar Rapids

The Cedar River in Cedar Rapids, which runs roughly from the northwest of the city to its southeast, crested at 31.12 feet on June 12, 2008, exceeding the 500-year floodplain area (FEMA P-765). The city experienced a total of $5.4 billion in damages and economic losses (FEMA P-765). The flooding affected an estimated 10 square miles (2589 hectares, or 14% of the city), including 1126 city blocks, nearly 5400 homes, over 800 commercial and government buildings, and displaced an estimated 18,000 people.[4] The areas near the river experienced the worst impacts, including the downtown area on the east side of river and a largely residential area along the west side of the river.

In the aftermath of the 2008 flood, the city developed the *Framework for Reinvestment and Revitalization*, outlining a vision for Cedar Rapids as a "vibrant urban hometown – a beacon for people and businesses invested in building a greater community for the next generation."[5] At the core of the *Framework* was an extensive Flood Management System, envisioned to protect a stretch of 7.5 miles along the Cedar River[6] and projected to take 20 years to complete, at an estimated cost of $375 million. Nearly ten years after the flood, the cost was estimated to be $550 million.[7] The key components of the Flood Management System include levees, permanent and removable walls, gates, and pump stations. The city also engaged in a land acquisition program, funded by federal grants, to protect land prone to flooding, largely on the west side of the river (Tate, et al. 2016).

In addition, the Framework emphasized "the creation of Sustainable Neighborhoods," resulting in a Neighborhood Reinvestment Plan approved by the City Council on May 13, 2009. The Neighborhood Reinvestment Plan emphasized neighborhood revitalization as another key component in addition to the flood-control infrastructure.[8] The revitalization focused on ten neighborhoods, including Downtown Cedar Rapids, as well as the adjacent New Bohemia (NewBo) neighborhood and the historic Czech Village neighborhood across the river. Today, Downtown Cedar Rapids, NewBo, and Czech Village have become vibrant neighborhoods, attracting young professionals, entrepreneurs, and artists.

---

[3] Note that Bond et al. (2017) define the resilience dividend as "the difference in net benefits from a project developed with a resilience lens versus one that is not," which is much broader than the definition used in this paper (Fung and Helgeson 2017).
[4] See the City of Cedar Rapids, Flood of 2008 Facts and Statistics: http://www.cedar-rapids.org/discover_cedar_rapids/flood_of_2008/2008_flood_facts.php.
[5] See City of Cedar Rapids, Flood Recover Planning: http://www.cedar-rapids.org/local_government/departments_a_-_f/community_development/flood_recovery_planning.php.
[6] See City of Cedar Rapids, Flood Management System: http://www.cedar-rapids.org/discover_cedar_rapids/flood_of_2008/flood_management_system.php.
[7] See The Gazette, July 5, 2018: https://www.thegazette.com/subject/news/government/cedar-rapids-flood-protection-funding-approved-army-corp-joni-ernst-iowa-2008-flood-20180706
[8] See City of Cedar Rapids, Neighborhood Reinvestment Action Plans: http://www.cedar-rapids.org/local_government/departments_a_-_f/community_development/neighborhood_reinvestment_action_plans.php.

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

Figure 1 presents a map of the city of Cedar Rapids, highlighting the three neighborhoods of Downtown Cedar Rapids, NewBo, and Czech Village, which have been particular targets for commercial and residential development to attract a younger, more dynamic work force. Due to their size relative to the whole economy, the CGE model combines Downtown Cedar Rapids, NewBo, and Czech Village into a single spatial unit. For simplicity, the combined spatial unit is called "Downtown."

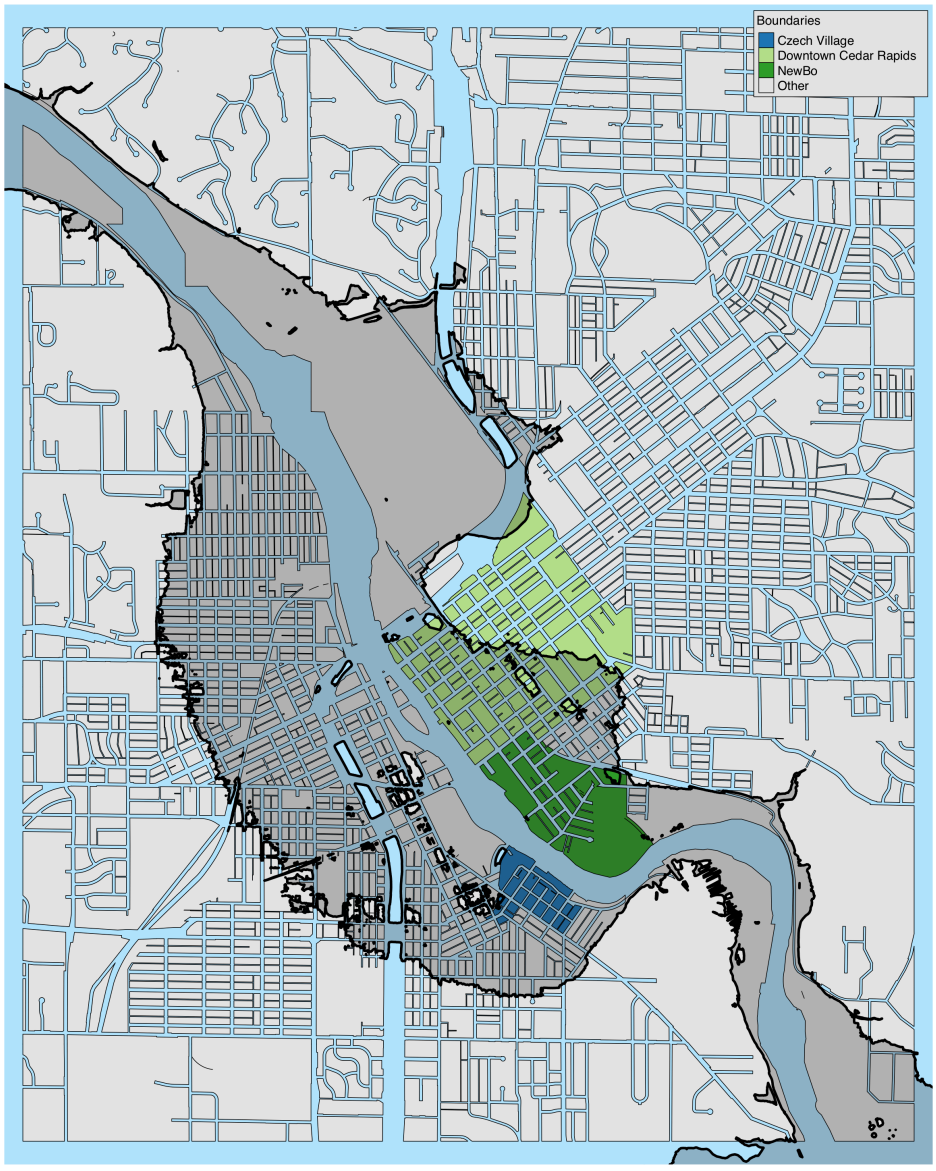


Figure 1. Detail of the "Downtown" area neighborhoods and extent of the 2008 flood (dark grey shading with black boundary). Map created using city of Cedar Rapids shapefiles.

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

It should be noted that 2008 ushered in another major catastrophe, one that affected the entire country. The Great Recession, which is officially recognized as beginning on December 2007 and ending on June 2009, saw large declines in Gross Domestic Product (GDP), home prices, and stock markets, while the national unemployment rate rose to 10 percent by October 2009.[9] Such added downward pressure on the local economy makes the path Cedar Rapids took seem even more impressive. While it is impossible to disentangle the effects of the recession from the effects of the flood, the impacts of reinvestment are expected to move in the opposite direction. Thus, they may be understated in our results.

3. The CGE approach to quantifying the resilience dividend

3.1. The CGE approach in this paper

This paper uses a comparative-static spatial CGE modeling approach. Comparative-static approaches compare two alternative equilibria in order to assess the impact of shocks to the economy.[10] The process of adjustment from the old (status quo) equilibrium to the new equilibrium is not explicitly represented in such a model, as the temporal element of a CGE model is not well defined.[11] Nevertheless, the difference between the status quo and the new equilibrium is attributed solely to the shock and thus, impacts of the shock are quantified through changes to prices and quantities.

Spatial CGE (SCGE) models allow for a geographic distribution of the impacts from shocks to an economy. Thus, SCGE models are a natural fit for exploring the distributive effects (in particular, the resilience dividend) of resilience planning associate with large-scale shocks across a community.

This paper uses annual economic data to build *two* SCGE models of Cedar Rapids: one based on the 2007 data and the other based on the 2015 data. Each model represents a status quo (pre-shock) equilibrium. In other words, the two models are "snapshots" of the pre-2008 economy and of the post-2008 economy, respectively. The snapshots may be thought of as alternative states of the world, providing plausible counterfactuals for quantifying the resilience dividend. In particular, if each snapshot responds to the same shock in different ways, the differences can be attributed to investments in resilience. In Sec. 6, we consider a shock that increases total factor productivity in each economy.

3.2. Limitations of the CGE approach to quantifying resilience dividend

The ultimate goal of the proposed SCGE modeling method is to quantify the resilience dividend. A CGE model provides distributional impacts of shocks, policy changes, and the current status of the region. Distributional impacts allow the analyst to understand not only the overarching net impacts, but to whom and where those impacts fall and are distributed. Large economic effects will be easily discerned, and the impacts can be selected to see how different scenarios may have played out in the region. Any effects of resilience actions that have co-benefits can be modeled to identify how those co-benefits manifest themselves throughout the economy. Thus, the resilience dividend can be quantified as a grand total, as well as determining who gets these benefits and where they go spatially. On the other hand, CGE models may not capture the entirety of the resilience dividend in many cases. Non-market benefits that never

[9] Federal Reserve History, The Great Recession: https://www.federalreservehistory.org/essays/great_recession_of_200709.
[10] In contrast, dynamic CGE models explicitly trace each model variable through time in order to capture the path to equilibrium (Pereira and Shoven 1988). Helgeson et al. (2017) discuss the potential difficulties with this approach.
[11] However, it is possible to distinguish between short-run and long-run equilibria (e.g., looking at whether capital stocks are allowed to adjust in a given run of the model).

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

actually materialize as real cash flows are not necessarily captured. Minor impacts may also be lost as the overall economic conditions may overwhelm them.

4. Data for the CGE approach

The primary objective of CGE data collection is to develop a social accounting matrix (SAM). A SAM quantifies all cash flows between pertinent actors within an economy (Hirway, Saluja and Yadav 2008). Once the SAM is constructed, it must be "balanced:" payments made by each component of the SAM should exactly equal payments received by each component of the SAM. A balanced SAM is the main input to the CGE model and represents the status quo for the economy.

The model is "calibrated" when the CGE model equations solve for an equilibrium (set of prices and quantities that clear markets) that exactly reproduces the status quo. Once the CGE model is calibrated, it is ready for simulated shocks. Shocks are applied to exogenous parameters of the CGE model (e.g., total factor productivity) and a new equilibrium is found. Comparative-static analysis compares the "status quo" equilibrium and the new equilibrium.

The following primary data sources are used to build the Cedar Rapids SAMs, based on the method for constructing a spatial SAM and CGE model developed in Cutler et al. (2017).

4.1. Quarterly Census of Employment and Wages (QCEW) data

We obtained establishment-level Quarterly Census of Employment and Wages (QCEW) data from 2006 to 2016 for Linn County, IA, of which Cedar Rapids is the county seat. The QCEW reports the quarterly count of employment and wages for establishments and includes the establishment's industry, defined by its North American Industry and Classification System (NAICS) code, and street address. This data is an excellent source to determine wage payments, employment, and number of firms by industry.[12] The advantage of establishment-level data is that we can customize how the data is aggregated into productive sectors, as discussed in Section 4.7. Moreover, since many of the records in the QCEW data contain street addresses for the establishments, we break sectors out spatially (e.g., "Downtown" versus "Other") as discussed in Section 4.6.

4.2. Public Use Microdata Sample (PUMS) data

Public Use Microdata Sample (PUMS) data allows us to determine the distribution of workers and wage payments by sector, household, and labor group, where households and labor groups are defined by household income and earnings, respectively. PUMS data is collected by the U.S. Census Bureau and reported at various intervals. PUMS relies on the use of American Community Survey (ACS) data. Unlike the decennial census, ACS surveys are conducted annually. Roughly one in thirty-eight households are invited to take the survey every year.[13] The data collected in the ACS is very similar to the data collected during the decennial census. The household income distribution can be obtained from this data set.

[12] Because the data contains commercially identifiable information (CII) and, potentially, personally identifiable information (PII), such QCEW data is not publicly available. The QCEW data we obtained is collected by the Iowa Division of Labor for the BLS.

[13] See U.S. Census Bureau, "American Community Survey: how it works for your community." https://www.census.gov/content/dam/Census/programs-surveys/acs/about/2017%20How%20the%20ACS%20Works%20for%20Your%20Community_508.pdf.

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

We use PUMS data for the Linn County-Cedar Rapids Public Use Microdata Area (PUMA).[14] The Linn County-Cedar Rapids PUMA coincides with Linn County, and includes the city of Cedar Rapids as well as cities such as Hiawatha and Marion whose economies are intertwined with Cedar Rapids. Sectors are aggregated using NAICS codes. As noted below in Section 4.8, the NAICS codes for QCEW and PUMS data do not map one-to-one, so an intermediate step is necessary to map each NAICS code to one of our custom-defined sectors.

4.3. County and City Assessor data

We collected data from both the Linn County Assessor and the City of Cedar Rapids Assessor for 2006 to 2016. The tax assessment data includes parcel-level assessed values for land and improvements (i.e., the value of a building), which in turn are used to estimate land and capital values, respectively, in the SAM. In addition, by estimating household expenditures on housing services, such as rent and interest paid on mortgages,[15] we can estimate the value of housing services in the economy from assessments. Finally, the assessor data includes street addresses for each parcel. We can geocode the parcels in order to match capital and land values for non-residential buildings to the establishment-level QCEW data, as discussed further in Section 4.6.

4.4. City Budgets and the Comprehensive Annual Financial Report (CAFR)

A Comprehensive Annual Financial Report (CAFR) contains details of the financial state of a given governmental entity such as a state or municipality. CAFRs for the city of Cedar Rapids are publicly available online by fiscal year.[16] The CAFRs are useful resources for the determination of local government tax revenue, expenditures, and employment. The Cedar Rapids CAFR provides the information necessary to decompose employment and expenditures into constituent government "industries" (e.g., education, public health, and public safety). This information is critical to properly size and disaggregate the government sector within the CGE model.

4.5. Bureau of Economic Analysis (BEA) data

Bureau of Economic Analysis (BEA) data is vital in building the SAM. The BEA provides the Input-Output Accounts data needed to determine Input-Output (I-O) coefficients for the SAM and the values required to develop the relationship between investment and the stock of capital.[17]

The I-O data is generally taken at the national level and, in its raw form, gives the raw dollar amounts of input from each industry and the total output from each industry. These values can be used to determine I-O coefficients, which represent how much input each industry requires from every other industry in order to produce a dollar's worth of output. I-O coefficients define the flow of money between industries, and thus the linkages between industries necessary for the CGE model to determine how impacts on one industry flow to another.

The data for the investment capital linkage (CAPCOM) matrix comes from the BEA "Capital Flow" data. This data tracks the investment in new structures, equipment, and software by using industries. In essence, it measures how many commodities a specific industry purchases for investment from another

[14] PUMAs are "statistical geographic areas defined for the dissemination of Public Use Microdata Sample (PUMS) data" (https://www.census.gov/geo/reference/puma.html).
[15] We use the BLS Consumer Expenditure Survey category "Shelter" for these expenditures. See https://www.bls.gov/cex/csxgloss.htm#housing for more details.
[16] See http://www.cedar-rapids.org/local_government/departments_a_-_f/finance/cafr.php.
[17] See https://www.bea.gov/industry/input-output-accounts-data for more information.

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

industry. Like the I-O data, the CAPCOM tracks the interdependencies between industries; however, it focuses on new investments instead of required input. The raw data is taken from the I-O commodity categories (as opposed to the National Income and Product Account categories), which are in terms of producers' prices.

Other BEA data used for our model include estimates of employment and income, which are available at the county level. The BEA series "Personal Income and Employment by Major Component" provide estimates of income, population, and employment. Personal income is broken down by source (e.g., wages and salaries, contributions to government social insurance), while employment is broken down by wage earners and proprietors. The BEA series "Total Full-time and Part-time Employment by NAICS Industry" breaks employment down further by high-level (i.e., two-digit) NAICS code. These data sets offer a high-level, order of magnitude check on the PUMS and QCEW data on employment and wages.

4.6. Geographic data

We obtained parcel-level and boundary geographic information systems (GIS) data for the city of Cedar Rapids.[18] We use this data to visually assign parcels to the downtown neighborhoods of interest (Downtown Cedar Rapids, NewBo, and Czech Village). Any parcels outside of this area are assigned to the rest of the economy and we label them as "Other." Once we define the neighborhoods spatially, we can match the geocoded QCEW and assessor data to the neighborhoods in order determine which neighborhood an establishment or parcel belongs to.

4.7. Informal data from community leadership and agencies

The community itself proved to be an invaluable source of information. Conversations with the City Manager's Office, Cedar Rapids Economic Development, and Go Cedar Rapids (the tourism office) illuminated priorities with respect to both the immediate response to the 2008 floods, as well as short- and long-term recovery efforts and community goals. In particular, while we initially focused on the land acquisition program, conversations with local officials quickly revealed that revitalizing the downtown area was a key component of rebuilding after the 2008 floods.

Conversations with community officials also provided perspective on local economic trends and goals, both irrespective of the potential disaster and specific to the disaster occurrence. This informed how we defined the productive sectors for the model, by allowing us to focus on sectors the community itself identified as important. This was particularly helpful in identifying the sectors that are important in the Downtown area, as discussed in Section 5.1.

4.8. Combining the various data sets

The use of such varied sources of data can create challenges when combining them for the SAM (Helgeson, et al. 2017). One example of this complication is attempting to derive the I-O and CAPCOM data at the PUMS sector. The BEA and PUMS data sets are both based on NAICS codes; however, they aggregate those NAICS codes into larger industry categories that do not match one-to-one with each other. If industries are defined coarsely, this is not necessarily an issue. If the manufacturing industry data is disaggregated, as in our model, then there is no guarantee that each PUMS industry code will have a corresponding BEA code, or codes, that match in terms of NAICS codes covered. In this case, a fuzzy

[18] See City of Cedar Rapids GIS Division: http://www.cedar-rapids.org/local_government/departments_g_-_v/information_technology/available_gis_data.php.

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

match is required which will possibly lead to a NAICS code from a sector not in a specific PUMS industry code being in the I-O table for that PUMS industry code due to the inconsistency.

5. Modeling the Cedar Rapids economy

Cedar Rapids is the largest city in, and the county seat of, Linn County, Iowa. Cedar Rapids is an integral part of a regional economy that includes the neighboring cities of Marion, Hiawatha, Mount Vernon, and Robins, which together comprise the five most populous cities in Linn County. Given the close economic relationships between Cedar Rapids and the other cities in Linn County, this paper models the regional economy of Cedar Rapids as encompassing Linn County.

5.1. Important sectors

Figure 1 presents the largest employers in the city of Cedar Rapids in 2015, and their relative share of county employment for both 2007 and 2015. Note that two hospitals (St. Luke's and Mercy), the Cedar Rapids Community School District, and the city itself, are some of the largest employers in the city.

Employer	2007		2015	
	Employees	Percentage of Total County Employment	Employees	Percentage of Total County Employment
Rockwell Collins Inc.	9000	5.41%	8700	4.95%
Transmerica / Aegon	3500	2.10%	3800	2.16%
St. Luke's Hospital	2800	1.68%	2979	1.69%
Cedar Rapids Community School District	2900	1.74%	2879	1.64%
Nordstrom Direct	2862	1.79%	2150	1.22%
Mercy Medical Center	2498	1.50%	2140	1.22%
City of Cedar Rapids	1493	0.90%	1309	0.74%
Four Oaks			1100	0.63%
Quaker Foods and Snacks	1100	0.66%	1018	0.58%

Table 1. Principal employers in Cedar Rapids. Source: Cedar Rapids CAFRs, FY2015 and FY2007.

The city itself identified five "target industries" in developing a strategic economic development plan in 2014:[19]

- Life Sciences
- Logistics and Distribution
- Food Sciences and Processing
- Entrepreneurial Business Services, and
- Finance, Insurance, and Real Estate.

Based on the city's self-identified target industries, as well as on the industries that are important to the downtown area of Cedar Rapids, we defined the Cedar Rapids regional economy's productive sectors as shown in Table 1 and Table 2.[20] The corresponding two-digit NAICS codes and high-level NACIS industry names are also shown. The data used for the CGE model includes 6-digit NAICS codes, which

[19] See City of Cedar Rapids, Economic Development: http://www.cedar-rapids.org/local_government/departments_a_-_f/community_development/economic_development_services.phd

[20] Some of these industries were emphasized during conversations with City officials. See Sec. 4.8.

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

provide a much finer level of industry detail. As discussed in Section 4, we define these sectors by aggregating establishment-level employment and wage data.

Table 2. Sectors chosen for the Cedar Rapids CGE models that are not present in downtown Cedar Rapids.

Sector name	NAICS code	NAICS industry title
Electronics manufacturing	33	Manufacturing
Food processing	31	Manufacturing
Paper manufacturing	32	Manufacturing
All other manufacturing	31-33	Manufacturing
Construction	23	Construction
Transportation	48-49	Transportation and Warehousing
Online services	45 49	Retail Trade Transportation and Warehousing
Education	61	Educational Services
Health care	62	Health Care and Social Assistance
Wholesale trade	42	Wholesale Trade
Information	51	Information
Agriculture and mining	11 21	Agriculture, Forestry, Fishing and Hunting Mining
Utilities	22	Utilities

Note that manufacturing is broken down into four separate sectors: Electronics, Food, Paper, and All other manufacturing. Another key sector, Online services, includes retail and logistics, reflecting the importance of online retailers (e.g., Nordstrom Direct in Fig. 1). Non retail-oriented logistics are included in the Transportation sector. Moreover, Agriculture and mining are combined into a single sector due to their relatively small contribution to the local economy.

Table 3. Sectors chosen for the Cedar Rapids CGE models that are present within and outside downtown.

Sector name	NAICS code	NAICS industry title
Financial and insurance services	52	Finance and Insurance
Real estate services	53	Real Estate Rental and Leasing
Professional business services	54 55	Professional, Scientific, and Technical Services Management of Companies and Enterprises
Services	56 81	Administrative and Support and Waste Management and Remediation Services Other Services (except Public Administration)
Arts and entertainment	71	Arts, Entertainment, and Recreation
Accommodation	72	Accommodation and Food Services
Restaurants	72	Accommodation and Food Services
Retail	44-45	Retail Trade

The sectors in Table 2 represent the core sectors in found both within downtown Cedar Rapids and throughout the rest of the economy. In the CGE model, these sectors are identified spatially by the location of the firm (i.e., whether or not the firm is located in downtown Cedar Rapids). Professional business services (PBS) covers two distinct NAICS industries and reflects one of the city's self-identified target industries. Finally, Accommodation and Restaurants are separated out of NAICS code 72.

5.2. Summary Statistics: 2007 and 2015 Snapshots

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

This section presents select aggregate economic statistics that provide a snapshot of the Cedar Rapids regional economy in each of 2007 and 2015. Note that these are status quo outcomes in each model, rather than the results of a shock. In other words, these are "snapshots" of the pre-2008 economy and of the post-2008 economy.

Table 3 presents land and capital values, as well as total acres, for 2007 and 2015. Growth in the combined Downtown area was more pronounced: total capital values grew approximately 144% in the Downtown, while capital values in the rest of the economy grew 83%. Land values, on the other hand, only grew about 4.05% in the Downtown area. In contrast, growth in land values for the rest of the economy was about 37%.

Table 4. Land and capital values (in millions of dollars) and total acres for the Downtown area (Downtown Cedar Rapids, NewBo, and Czech Village) and the rest of the regional economy by year, based on County Assessor data for Linn County, IA.

District	Year	Land	Capital	Acres (Hectares)
Downtown	2007	42.74	384.58	197.67 (80.00)
Downtown	2015	44.78	824.64	477.27 (193.15)
Other	2007	2,739.99	10,345.41	370,559.25 (149,960.01)
Other	2015	3,755.79	18,948.89	498,265.87 (201,641.04)

During this period, the Downtown area grew in size by a factor of about 2.5, while the area of the rest of the economy only grew about 34%. Together with growth in capital, the growth in acreage reflects significant investment in developing Downtown relative to the rest of the economy.

Table 4 presents total employment and wages paid per worker for each year. While total employment in the Downtown area only grew by about 2.1% (compared to about 5.4% in the rest of the economy), wage per worker grew by 26.5% Downtown (compared to about 22.7% in the rest of the economy). Thus, while employment growth Downtown does not reflect the trend in capital and land area, wage per worker does appear to be growing slightly faster in Downtown.

Table 5. Employment (number of workers) and annual wage per worker (in dollars) for the Downtown area (Downtown Cedar Rapids, NewBo, and Czech Village) and the rest of the regional economy by year, based on the Quarterly Census of Employment and Wages (QCEW) for the state of Iowa.

District	Year	Employment	Wage per worker
Downtown	2007	5,801	11,244.89
Downtown	2015	5,924	14,230.58
Other	2007	115,080	10,556.87
Other	2015	121,296	12,951.05

6. Main results

6.1. Description of shocks

As a first step toward quantifying the resilience dividend, we compare how pre-2008 and post-2008 Cedar Rapids respond to a similar, non-disaster shock. A differential response to the same shock can be largely

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

attributed to investing in resilience, including revitalization of downtown. We consider a positive shock to the economy: total factor productivity (TFP) increases by 2%. The intuition for the shock is that the economy uniformly becomes more productive (e.g., due to new advances in technology). In this scenario, which economy is better situated to reap the benefits of increased productivity?

6.2. Impacts of TFP shocks on output and employment

This section presents the impacts of the TFP shock on two important macroeconomic indicators: output and employment. The results show that post-2008 Cedar Rapids experienced greater benefits from the TFP shock than pre-2008 Cedar Rapids.

The columns in Table 6 present output, defined as domestic supply, in 2007 and 2015 Cedar Rapids. The rows present output before and after the TFP shock, as well as the change in output from the shock. As shown in Table 6, grew 5.1% in 2015 Cedar Rapids, compared with 1.7% growth in 2007 Cedar Rapids. This amounts to 3.4% greater output growth from the TFP shock in 2015 than in 2007. The additional growth in 2015 is a co-benefit of investing in increased resilience and is thus part of the resilience dividend. To put the impact on output in context, recall that the resilient flood-control system is estimated to cost $550 million over ten years, while the growth in domestic supply alone is $648 million, which does not include other co-benefits to the economy such as employment growth.

Table 6. Output (domestic supply) in 2007 and 2015, both before (pre-shock) and after (post-shock) a TFP increase of 2%. The TFP shock leads to proportionately larger output growth in 2015. Output values are in millions of dollars.

	2007	**2015**
Pre-shock	9783.05	12037.12
Post-shock	9956.99	12685.63
Difference	173.94	648.51
Percent change	1.7%	5.1%
Resilience dividend		**3.4%**

Table 7 presents a similar picture for employment. While TFP leads to 0.61% growth in 2007 employment, 2015 employment growth is about twice as large at 1.2%. This amounts to 0.59% greater employment growth from the TFP shock in 2015 than in 2007, another co-benefit of investing in increased resilience.

Table 7. Employment in 2007 and 2015, both before (pre-shock) and after (post-shock) a TFP increase of 2%. The TFP shock leads to proportionately larger employment growth in 2015. Employment values are in total number of workers.

	2007	**2015**
Pre-shock	93903	120843
Post-shock	94483	122348
Difference	580	1505
Percent change	0.61%	1.2%
Resilience dividend		**0.59%**

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

6.3. Intuition for results

Our CGE results indicate that sector output in Cedar Rapids had a greater response to the TFP shock in 2015 than in 2007. The explanation of the differences lies in the calculation of the sector level estimates for total factor productivity (TFP) *before* the shock. Consider the following general production function:

$$Y_i = \delta_i f(L_{ji}, K_i, LA_i),$$

where Y is output, *i* is the index across commercial sectors, δ_i is the estimate of sector level TFP, L is employment, the index *j* indexes labor groups and K is capital, and LA is land. In the construction of the SAM, sector level output, Y_i, is calculated and we have collected data for L, K and LA. Therefore, we can solve for δ_i as

$$\delta_i = Y_i / f(L_{ji}, K_i, LA_i).$$

The initial values for δ_i have important implications when TFP shocks are simulated. The larger initial value of δ_i implies that a given positive shock to δ_i will result in a larger impact on sector level output. The differential impacts from the same shock reflect the relative difference of the initial δ_i's across the two time periods. Table 6 presents the difference of 2015 and 2007 estimates of δ_i.

Table 8. Differences in the calculated initial values for δ_i (2015 – 2007) by sector.

Sector name	**Difference in δ_i**
Electronics manufacturing	0.310
Food processing	0.324
Paper manufacturing	0.052
Other manufacturing	0.009
Construction	0.433
Transportation	-0.390
Online	0.090
Finance and insurance	1.418
Finance and insurance (Downtown)	1.422
Real estate	-0.144
Real estate (Downtown)	-0.214
Professional business services	0.139
Professional business services (Downtown)	0.210
Education	-0.152
Health	-0.046
Services	-0.114
Services (Downtown)	-0.033
Arts and entertainment	-0.075
Arts and entertainment (Downtown)	-0.012
Accommodation	-0.022
Accommodation (Downtown)	0.075
Restaurants	0.494
Restaurants (Downtown)	0.347
Information	0.323
Wholesale trade	0.165

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

Retail	0.486
Retail (Downtown)	0.477
AGMIN	-0.074
Utilities	-0.011

The vast majority of sectors experienced an increase in the estimated δ_i from 2007 to 2015. Without a doubt the change in the downtown area was paramount to other changes in Cedar Rapids during this period. As Table 6 indicates, the expansion of the downtown area for finance/insurance, professional business services, restaurants, accommodation, and retail all experienced an increase in δ_i. This is consistent with a denser allocation of commercial sectors in the downtown area and the resulting higher values for δ_i. It is worth pointing out that most of the estimates for δ_i also increased for the sectors located outside of the downtown area. For the same shock to each economy, the higher values for δ_i in 2015 will result in a larger increase in sector level output and total output for the 2015 model.

7. Conclusion and future work

This paper presents a spatial CGE approach to quantifying the resilience dividend. In particular, we build two snapshots of Cedar Rapids (pre-2008 and post-2008) that serve as counterfactuals of an economy with and without investments in increased resilience. By simulating the same shock to each snapshot, we can quantify how impacts differ in Cedar Rapids pre-resilience versus Cedar Rapids post-resilience. In particular, we find that the same increase in total factor productivity leads to an order of magnitude larger increases in employment and output in post-2008 Cedar Rapids than in pre-2008 Cedar Rapids. This difference is the resilience dividend.

In future work, we will consider how pre-2008 and post-2008 Cedar Rapids respond to a wide range of positive and negative shocks to the economy. We also explore how each economy responds to a simulated natural disaster, which captures the *direct* impact of investing in increased resilience.

Acknowledgements

We are grateful to Sandi Fowler and Jeff Pomeranz (Cedar Rapids City Manager's Office), Donna Burkett and James Morris (Iowa Workforce Development), Mark Castenson (Linn County Assessor's Office) for invaluable information and insight, as well as David Butry (NIST), Douglas Thomas (NIST), and Chris Clavin (NIST) for their comments.

Fung, Juan; Helgeson, Jennifer; O'Fallon, Cheyney; Webb, David; Cutler, Harvey.
"Quantifying Macroeconomic Resilience Dividends in Cedar Rapids."
Paper presented at The 27th International Input-Output Association Conference, Glasgow, United Kingdom. June 30, 2019 - July 5, 2019.

Bibliography

Bond, Craig, Aaron Strong, Nicholas Burger, and Sarah Weilant. 2017. *Guide to the Resilience Dividend Valuation Model.* Santa Monica, CA: RAND Corporation.

Cutler, Harvey, J Davis, Y Hu, K Kakpo, S McKee, M Shields, and S Zahran. 2017. *Developing a Methodology to Build Spatial SAMs and CGE Models.* Fort Collins, CO: Colorado State University.

Cutler, Harvey, Martin Shields, Daniele Tavani, and Sammy Zahran. 2016. "Integrating engineering outputs from natural disaster models into a dynamic spatial computable general equilibrium model of Centerville." *Sustainable and Resilient Infrastructure* 1: 169-187.

Federal Emergency Management Agency. 2008. "Iowa Severe Storms, Tornadoes, and Flooding." FEMA DR-1763, Washington, DC.

Federal Emergency Management Agency. 2009. "Midwest Floods of 2009 in Iowa and WIsconsin: Buildng Performance Observations, Recommendations, and Technical Guidance." FEMA P-765, Washington, DC.

Fung, Juan F, and Jennifer F Helgeson. 2017. *Defining the Resilience Dividend.* NIST Technical Note 1959, https://dx.doi.org/10.6028/NIST.TN.1959.

Helgeson, Jennifer, Juan Fung, Cheyney O'Fallon, David Webb, and Harvey Cutler. 2017. "Identifying and Quantifying the Resilience Dividend using Computable General Equilibrium Models: A Methodological Overview." *Proceedings of the 2nd International Workshop on Modelling of Physical, Economic and Social Systems for Resilience Assessment.* Luxembourg: European Union. 191-207.

Hirway, Indira, MR Saluja, and Bhupesh Yadav. 2008. *The impact of public employment strategies on gender equality and pro-poor economic development.* Research Project No. 34, New York: The Levy Economics Institute of Bard College.

Mechler, Reinhard, Junko Mochizuki, and Stefan Hochrainer-Stigler. 2016. *Disaster Risk Management and Fiscal Policy: Narratives, Tools, and Evidence Associated with Assessing Fiscal Risk and Building Resilience.* Policy Research Working Paper WPS 7635, Washington, DC: World Bank Group.

Pereira, Alfredo M, and John B Shoven. 1988. "Survey of dynamic computable general equilibrium models for tax policy evaluation." *Journal of Policy Modeling* 10: 401-436.

Rodin, Judith. 2014. *The Resilience Dividend: Being Strong in a World Where Things Go Wrong.* New York: PublicAffairs.

Tanner, Thomas, Swenja Surminski, Emily Wilkinson, Rrobert Reid, Jun Rentschler, and Sumati Rajput. 2016. *The Triple Dividend of Resilience: Realising development goals through the multiple benefits of disaster risk management.* London: Gloabl Facility for Disaster Reduction and Recovery (GFDRR) at the World Bank and Overseas Development Institute (ODI).

Tate, Eric, Aaron Strong, Travis Kraus, and Haoyi Xiong. 2016. "Flood recovery and property acquisition in Cedar Rapids, Iowa." *Natural Hazards* 80 (3): 2055-2079.

Measurement of Drag Coefficients through Vegetation Canopy

Ryan Falkenstein-Smith and Kevin McGrattan

National Institute of Standards and Technology
Gaithersburg, Maryland, USA

1 Introduction

The Fire Research Division of the National Institute of Standards and Technology (NIST) has developed several numerical models to predict the behavior of fires within buildings. One of the models, a computational fluid dynamics (CFD) code called the Fire Dynamics Simulator (FDS)[1], has been extended to model fires in the wildland-urban interface (WUI). One crucial component of this type of modeling is the proper treatment of wind-driven flow through vegetation. The objective of the experiments described in this work is to measure the drag coefficient of vegetation for an empirical sub-model appropriate for CFD.

Measurements of this type have been performed by other researchers[2–6], most of whom used wind tunnels of various sizes. In most cases, a single plant or small tree was positioned within the tunnel and the resistance force measured. However, such a measurement is not readily applicable to a CFD model which does not necessarily consider the tree as a whole but rather as a volume occupied by subgrid-scale objects that decrease the momentum of the gases flowing through. Some plants might be smaller than a characteristic grid cell, and some trees might be larger, but in either case, these objects are just momentum sinks within individual grid cells that require some drag coefficient that is appropriate to the local conditions.

2 Model Development

Consider a volume filled with a random collection of vegetative elements, like pine needles or leaves, as shown in Fig. 1. This volume can be regarded as a single grid cell in a CFD model for which the computational domain may span hundreds to thousands of meters. At a given instant in the numerical simulation, this grid cell would have, at the very least, an average flow speed, U, and gas density, ρ. The vegetation within the cell is typically modeled as a collection of subgrid-scale Lagrangian particles whose mass, size, and shape are characterized by a handful of parameters that can be determined with field measurements. These particles exert a force per unit volume given by:

$$F = \frac{N}{V_\mathrm{c}} \frac{\rho}{2} C_\mathrm{d} A_\mathrm{p} U^2 \tag{1}$$

where N is the number of elements, V_c is the volume of the grid cell, A_p is the projected area of a single element, and C_d is a drag coefficient. Similar configurations have already been been adapted in numerical investigations[7;8]. A more convenient way to describe the vegetation is by specifying the surface to volume ratio of each element, σ, the volume (packing) ratio of the collection of elements, β, and a shape factor, C_s, defined in this case as the ratio of the element's projected area to surface area. With this information, and the following relations:

$$C_\mathrm{s} = \frac{A_\mathrm{p}}{A_\mathrm{s}} \quad ; \quad \beta = \frac{N V_\mathrm{e}}{V_\mathrm{c}} \quad ; \quad \sigma = \frac{A_\mathrm{s}}{V_\mathrm{e}} \tag{2}$$

Falkenstein-Smith, Ryan; McGrattan, Kevin.
"Measurement of Drag Coefficients through Vegetation Canopy."
Paper presented at 15th International Conference and Exhibition on Fire Science and Engineering (Interflam 2019), London, United Kingdom.
July 1, 2019 - July 3, 2019.

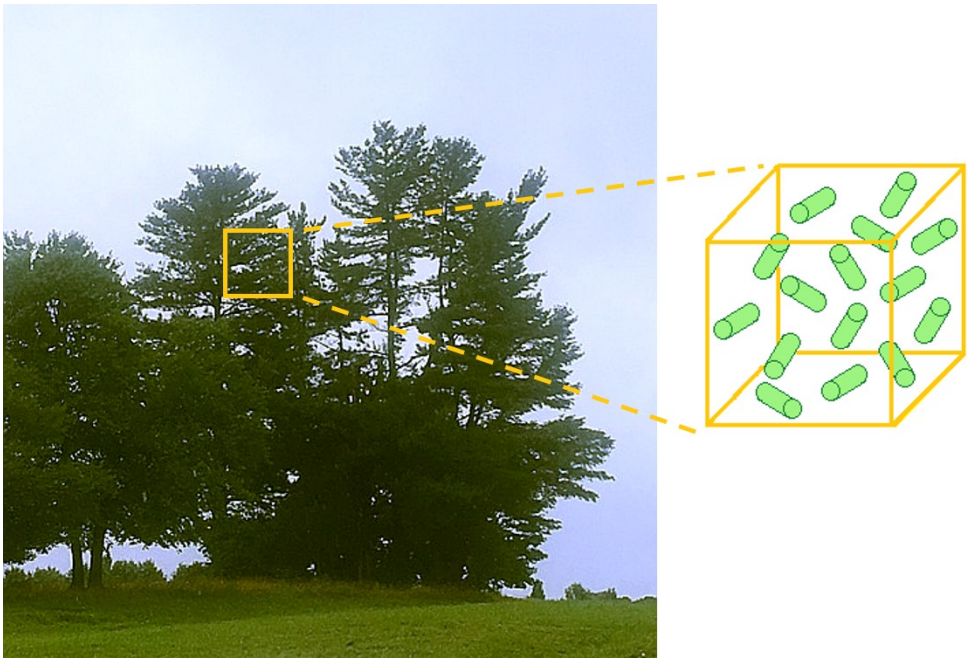


Fig. 1. Vegetation translation to multi-component model

where V_e is the volume of an element and A_s its surface area, we can convert the drag force expression in Eq. (1) to an equivalent form[9]:

$$F = \frac{\rho}{2} C_d \, C_s \, \beta \, \sigma \, U^2 \tag{3}$$

Some of the terms are difficult to measure, such as the shape factor and surface to volume ratio, but collectively these terms may be combined into a single parameter:

$$\kappa = C_s \, \beta \, \sigma \tag{4}$$

The parameter, κ, resembles an absorption coefficient* and can be determined by measuring the projected area of light, A, passing a given distance x through the vegetation. The decrease in the projected area of light is governed by the equation

$$\frac{dA}{dx} = -\kappa \, A \quad ; \quad A(x) = A(0) \, e^{-\kappa x} \tag{5}$$

The relative fraction of light passing through a distance of L is

$$W = \frac{A(L)}{A(0)} = e^{-\kappa L} \tag{6}$$

The parameter W is sometimes referred to as the "free-area coefficient" or "free-area fraction" in the literature.

In order to measure the drag coefficient, C_d, a section of length, L, of a small wind tunnel is to be filled with various amounts and types of vegetation and the pressure drop, ΔP, measured for

*Another way to express κ using the relations in Eq. (2) is NA_p/V_c, or in other words, the total projected area per unit volume. This parameter describes the absorption of non-scattering light by solid particles using the same geometric assumption for thermal radiation absorption.

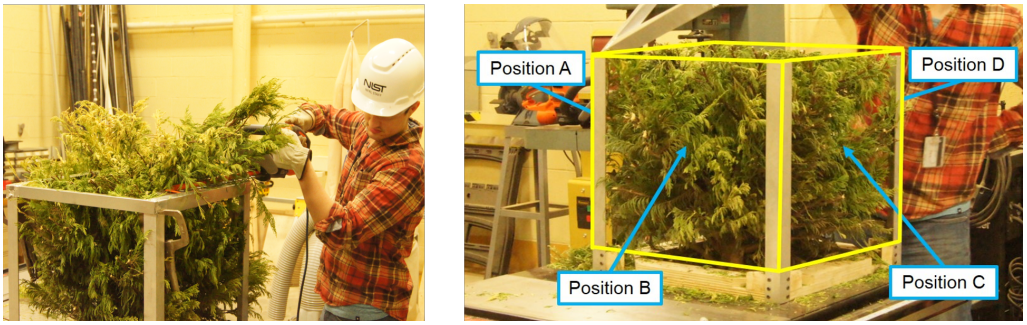


Fig. 2. Preparation of vegetation samples and designated orientation

an array of wind speeds, U. The value of κ shall be determined via black and white photography, and the drag coefficient extracted from the following form of the drag law derived above:

$$\frac{\Delta P}{L} = \frac{\rho}{2} C_{\mathrm{d}} \, \kappa \, U^2 \tag{7}$$

3 Description of Experiments

3.1 Sample Preparation

The vegetation chosen for this work was a Bakers Blue Spruce (*Picea pungens 'Bakeri'*), an Evergreen Distylium (Distylium 'PIIDIST-I'), a Gold Rider Leyland Cypress (*Cupressocyparis leylandii 'Gold Rider'*), a Kimberly Queen Fern (*Nephrolepis obliterata 'Kimberly Queen'*), a Blue Shag Eastern White Pine (*Pinus strobus 'Blue Shag'*), and a Robin Red Holly (*Ilex opaca*). Each sample was chosen based on its local availability. Leaf shapes were varied, including needle, elliptic, scale, and ovate.

The plant samples were cut into 0.5 m by 0.5 m by 0.5 m cubes using a guiding frame (left side of Fig. 2). The samples completely filled the cross section of the wind tunnel forcing the flow to move through the vegetation as opposed to around it. To easily distinguish the front, back, left, and right side of the cube-shaped vegetation, each side was designated Position A, B, C, or D (right side of Fig. 2). After its initial cut, image analysis, wind tunnel measurements, and water displacement testing were conducted in subsequent order. Image analysis and wind tunnel measurments were conducted for each position to obtain a collection of drag coefficients relative to different κ values. In some cases, samples were pruned and tested again. In the case of the Bakers Blue Spruce, Gold Rider Leyland Cypress, and Robin Red Holly, four prunings were made with the final one being the removal of all leaves.

3.2 Determining the Free-Area Coefficient via Photography

The free-area coefficient, W, was determined by placing each vegetation sample on a table located between a large white backdrop and a 0.5 m by 0.5 m cardboard frame, the same dimensions as the tunnel cross section (Fig. 3). For each sample cut and position, the projected area was photographed. All images were captured using a Nikon D5600 camera placed on a tripod located approximately 3.6 m away from the sample. The white backdrop was illuminated using a collection of incandescent and LED lights.

Falkenstein-Smith, Ryan; McGrattan, Kevin.
"Measurement of Drag Coefficients through Vegetation Canopy."
Paper presented at 15th International Conference and Exhibition on Fire Science and Engineering (Interflam 2019), London, United Kingdom.
July 1, 2019 - July 3, 2019.

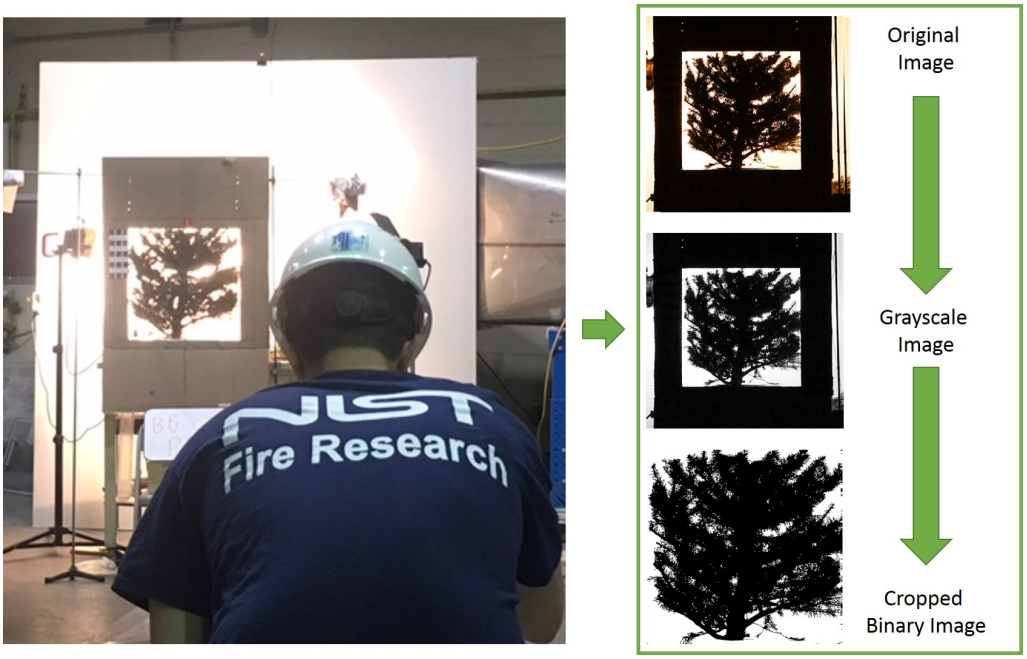


Fig. 3. Setup for photographing vegetation samples (left) and the post-processing procedure for analyzing images (right)

The images were processed using MATLAB's Image Processing Toolbox. Imported colored images were first converted into a grey scale and then a binary (black and white) image using a pre-set threshold level. The binary images were then cropped within the cardboard frame to eliminate non-vegetative substances and to evaluate the projected image of the vegetation exclusively. Once the projected image was obtained, a pixel count was conducted to determine the free-area coefficient of the vegetation, W. Once obtained, the free-area coefficient was used to calculate κ from Eq. (6).

3.3 Description of the Wind Tunnel

Pressure loss measurements were obtained in a wind tunnel test section with a cross-sectional area of 0.5 m by 0.5 m and a length of 2 m. An image and schematic diagram of the wind tunnel setup is shown in Fig. 4. The volume flow through the tunnel was measured upstream of the vegetation using a Rosemont 485 annubar[10]. The pressure drop across the vegetation was measured using an MKS Baratron Type 220D pressure transducer with a range of 0 to 133 Pa. The air flow was provided by a 0.91 m axial fan controlled by a variable frequency drive and monitored using the Annubar. Air density was calculated from pressure, temperature, and relative humidity readings of the testing facility. Each sample configuration was subjected to nine different fan speeds ranging from 0 to 88 % of the full-scale fan speed. The fan speed was not run at full scale due to the risk of exceeding the pressure transducer's pressure limitations. Data was sampled at 90 Hz for a 30 s period while maintaining a constant fan speed.

Once a set of measurements was taken at all fan speeds, the wind tunnel was shut off for approximately 5 min, and then the measurements were repeated. All measurements were repeated three times for each vegetation configuration. The variance homogeneity of the replicate

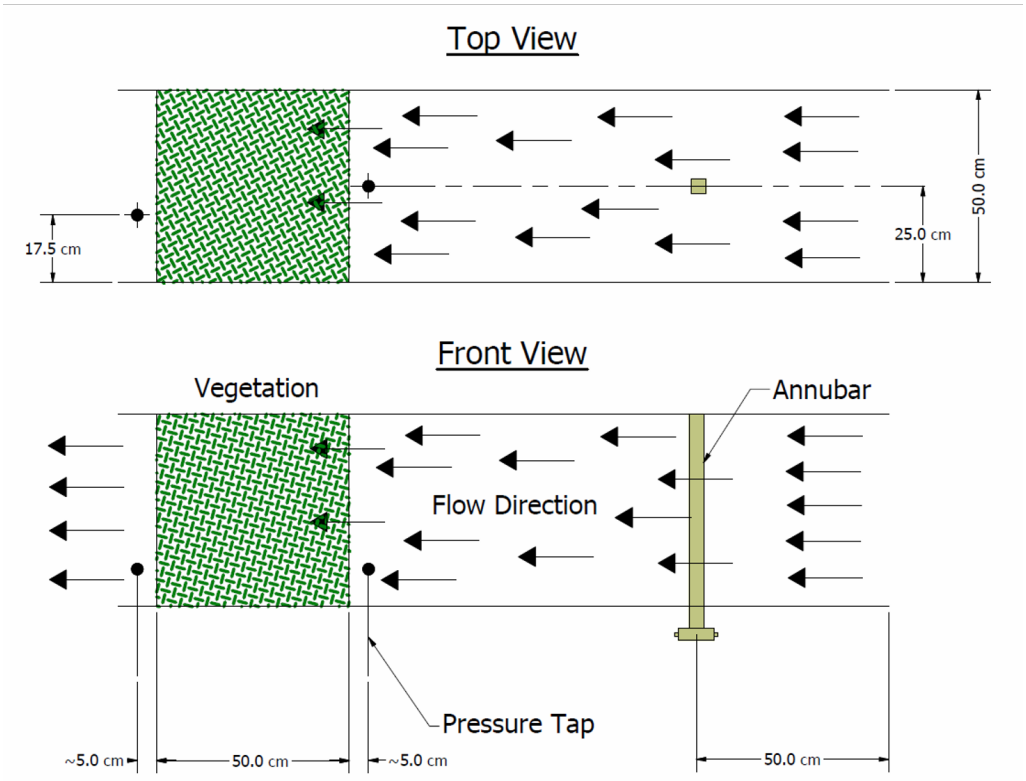


Fig. 4. Wind tunnel experimental setup with top and front schematic drawings

measurements was tested using Hartley's F_{max} test. If it was found that the data sets were homogenous, then the measurements were averaged.

3.4 Determining the Volume of Vegetation via Water Displacement

The volume of the vegetation was measured after a sample cut. The extracted vegetation was separated into branches and leaves and put into cloth mesh bags of known mass and volume, weighed, and submerged in a bucket. The displaced water flowed through a spout and into a

Fig. 5. Procedure of the water displacement test

beaker (Fig. 5). The measurement was repeated three times for each sample. The solid fraction, β, was calculated by dividing the average sample volume by the volume it occupied (0.5 m $\times$ 0.5 m $\times$ 0.5 m = 0.125 m^3).

4 Results

The key results of this work are the relationship between the absorption coefficient, κ, and the solid fraction, β, and the drag coefficient derived from the wind tunnel measurements.

4.1 Relationship between the Absorption Coefficient and Solid Fraction

Figure 6 presents the relationship between the averaged absorption coefficient, κ, and the solid fraction, β, for the sample configurations of the Bakers Blue Spruce, Evergreen Distylium, Gold Rider Leyland Cypress, and Robin Red Holly. The symbols indicate the measured values while the dotted lines represent a linear regression fit. Each line represents a particular type of vegetation that has been pruned, reducing both the volume fraction, β, the projected free-area coefficient, W, and the corresponding value of κ. There ought to be a linear relationship between κ and β if the shape factor, C_s, and surface to volume ratio, σ are constant, as shown in Eq. (4). However, this is not the case when the vegetative components are not uniform in size. Take, for example, the Robin Red Holly data shown in Fig. 6. As β decreases, κ should approach zero, as demonstrated by most samples. As the leaves of the Robin Red Holly were pruned, κ decreased significantly even though its volume fraction did not, owing to the fact the ratio of branch to leaf volume of the Robin Red Holly is substantially higher than the other plant species, as shown in Table 1. As a result, the free-surface area, W, decreases from the removal of leaves, thus reducing κ, while still maintaining a relatively consistent solid fraction due to the significant volume contribution of the branches.

Falkenstein-Smith, Ryan; McGrattan, Kevin.
"Measurement of Drag Coefficients through Vegetation Canopy."
Paper presented at 15th International Conference and Exhibition on Fire Science and Engineering (Interflam 2019), London, United Kingdom.
July 1, 2019 - July 3, 2019.

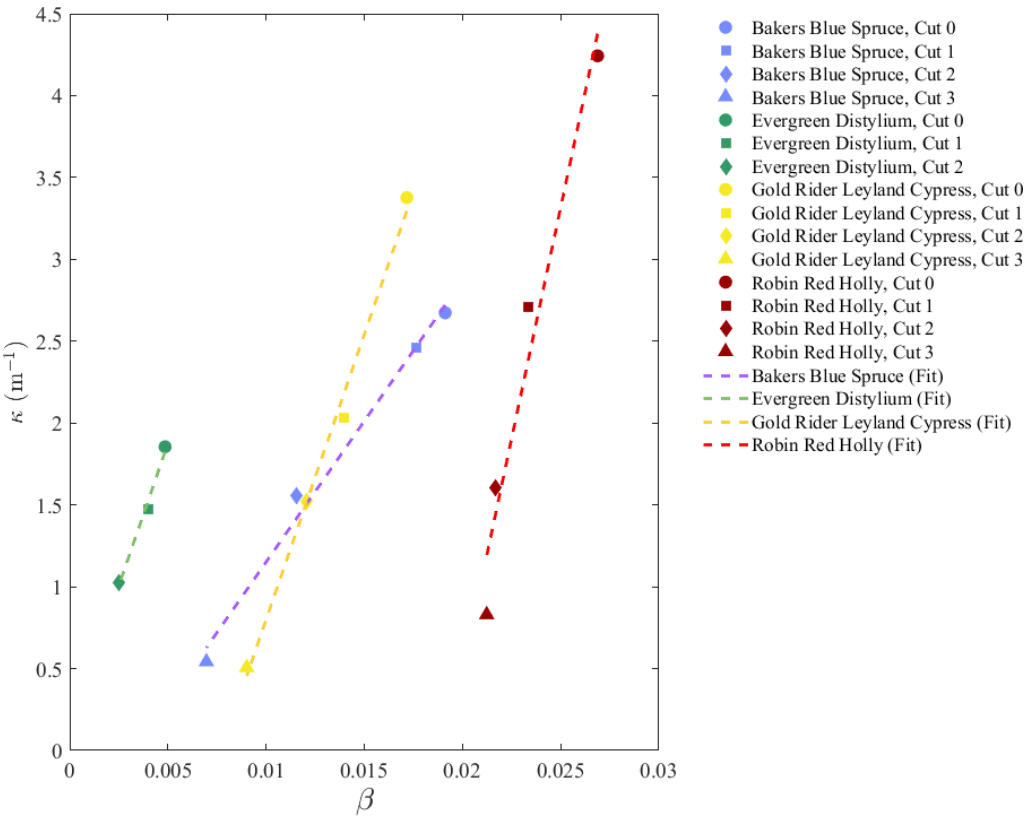


Fig. 6. Calculated absorption coefficient (κ) of vegetation sample configuration plotted against the corresponding solid fractions (β)

Table 1. Branch and leaf volume ratio of vegetation samples with mulitple cut iterations

Sample	β (%)	**Branch/Leaf Vol.**	**Sample**	β (%)	**Branch/Leaf Vol.**
Blue Spruce	1.9	1.1	Cypress	1.7	1.5
	1.8	1.3		1.4	2.2
	1.2	1.5		1.2	3.0
	0.7	N/A		0.9	N/A
Distylium	0.5	1.0	Red Holly	2.7	11
	0.4	1.4		2.3	16
	0.3	3.3		2.2	47
				2.1	N/A

4.2 Vegetation Canopy Drag Coefficients

The left plot of Fig. 7 displays the relationship between the freestream velocity and the pressure drop for each sample configuration. The results demonstrate the expected quadratic relationship. Replotting the data as shown in the center plot of Fig. 7 yields the drag coefficient for each sample configuration as determined by calculating the slope of each line of data points.

Falkenstein-Smith, Ryan; McGrattan, Kevin.
"Measurement of Drag Coefficients through Vegetation Canopy."
Paper presented at 15th International Conference and Exhibition on Fire Science and Engineering (Interflam 2019), London, United Kingdom.
July 1, 2019 - July 3, 2019.

No linear regression fitting was observed to have a coefficient of determination less than 0.98, indicating a close representation of the fitted regression line to the measured data. A summary of all 68 calculated drag coefficients and their respective uncertainties are presented in Table 2. The procedure for determining the drag coefficient uncertainty as shown in this table can be found in a previously published technical report[11].

The distribution of the measured drag coefficients for all sample configurations is shown in Fig. 8. The collection of sample configurations is divided into two groups based on leaf shape (i.e., narrow and broad). The narrow leaves group included the Bakers Blue Spruce, Blue Shag Eastern White Pine, and Gold Rider Leyland Cypress while the broad leaves group was comprised of the remaining species. The average drag coefficient of all sample configurations was determined to be 2.8 with an expanded uncertainty of 0.4.

To determine if the average drag coefficient depends on the type of vegetation a random effects one-way Analysis of Variance (ANOVA) was implemented on the drag coefficients of the different vegetation samples. The analysis yielded a significant variation ($F(5, 62) = 4.88$, $p = 7.97 \times 10^{-4}$) among the species[†]. A Tukey's test[12] was subsequently applied to determine if the species-specific average drag coefficients were significantly different from each other. The results showed one significant difference between the species' average drag coefficients: the Robin Red Holly and Gold-Rider Leyland Cypress. Despite the significant difference, the average drag coefficients of these two plant species are still within the uncertainty bound of the overall drag coefficient and therefore are not large enough to have a practical implication.

Further analysis was conducted to compare the two leaf shape groups. A one-way random effects model[13] for the measurements of the narrow leaves group assumes that the drag coefficients are normally distributed:

$$C_{\mathrm{d},ij} \sim N(m_1, {u_{ij}}^2 + {\sigma_1}^2),\ i = 1, ..., 3; j = 1, ..., n_i \tag{8}$$

where i denotes the plant species (1 for Bakers Blue Spruce, 2 for Blue Shag Eastern White Pine, and 3 for Gold Rider Leyland Cypress), and j denotes the specific configuration of the plant in the tunnel. The sample size for a given plant species is n_i. The value of u_{ij} is the standard uncertainty of the measured drag coefficient for a specific species and configuration. The parameters m_1 and σ_1 are the mean and standard deviation for the narrow leaves group, respectively. The drag measurements of the broad leaves group are modeled in a similar way:

$$C_{\mathrm{d},ij} \sim N(m_2, {u_{ij}}^2 + {\sigma_2}^2),\ i = 4, ..., 6; j = 1, ..., n_i \tag{9}$$

where $i = 4$ for Distylium, $i = 5$ for Fern, and $i = 6$ for Red Holly). The parameters m_2 and σ_2 are the mean and standard deviation for the broad leaves group, respectively.

Using a Bayesian statistical model[14] with non-informative priors for m_1, m_2, σ_1, and σ_2, we obtain via Markov Chain Monte Carlo implemented in OpenBUGS[15] the posterior means and standard uncertainties of the parameters. These are: m_1 is 3.0 with an expanded (95 %) uncertainty of 0.3, m_2 is 2.5 with an expanded (95 %) uncertainty of 0.3, and the 95 % uncertainty interval for the difference $m_1 - m_2$ is (0.052, 0.87). This may be interpreted as a rejection of a hypothesis test of H_0: $m_1 - m_2 = 0$ at a level of 5 %. Despite the differences in the average drag coefficients of both groups, they both lie within the uncertainty bound of the overall average drag coefficient, which suggests that mean drag coefficient obtained from all samples could be a reasonable approximation when applied as a consistent drag coefficient for vegetation canopies in CFD models.

[†]The F refers to the statistic obtained from the F-test conducted in the ANOVA, the 5 and 62 in brackets represent the degrees of freedom, and the 4.88 is the actual F statistic derived from the ANOVA. The p refers to the significance level determined from the F statistic and the 7.97×10^{-4} is the actual p-value which was determined to be less than the chosen confidence level of 0.05, indicating a significant difference between the mean drag coefficients of the samples.

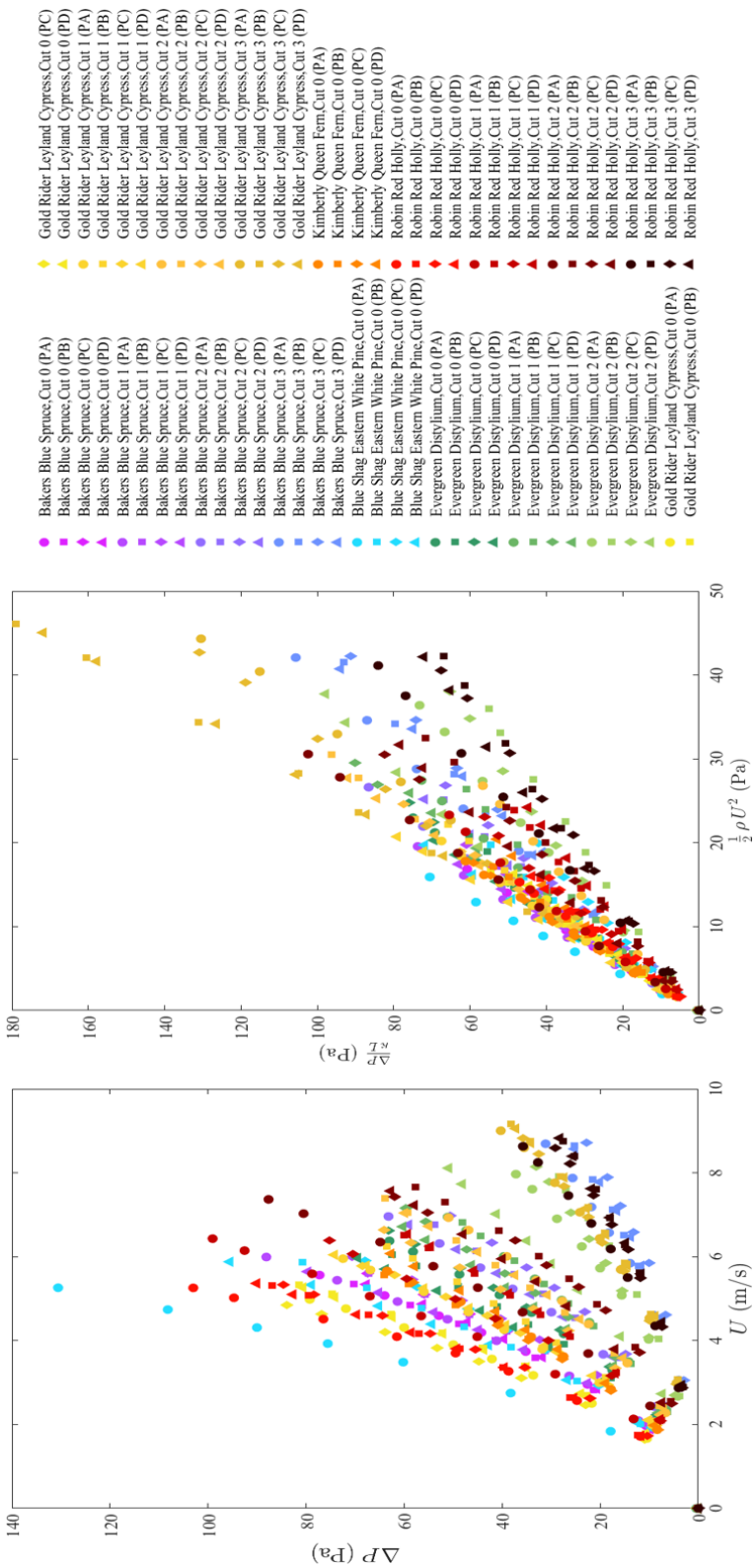


Fig. 7. Differential pressure measurments versus freestream velocity (left) and differential pressure over κ L versus dynamic pressure (center)

Falkenstein-Smith, Ryan; McGrattan, Kevin.
"Measurement of Drag Coefficients through Vegetation Canopy."
Paper presented at 15th International Conference and Exhibition on Fire Science and Engineering (Interflam 2019), London, United Kingdom.
July 1, 2019 - July 3, 2019.

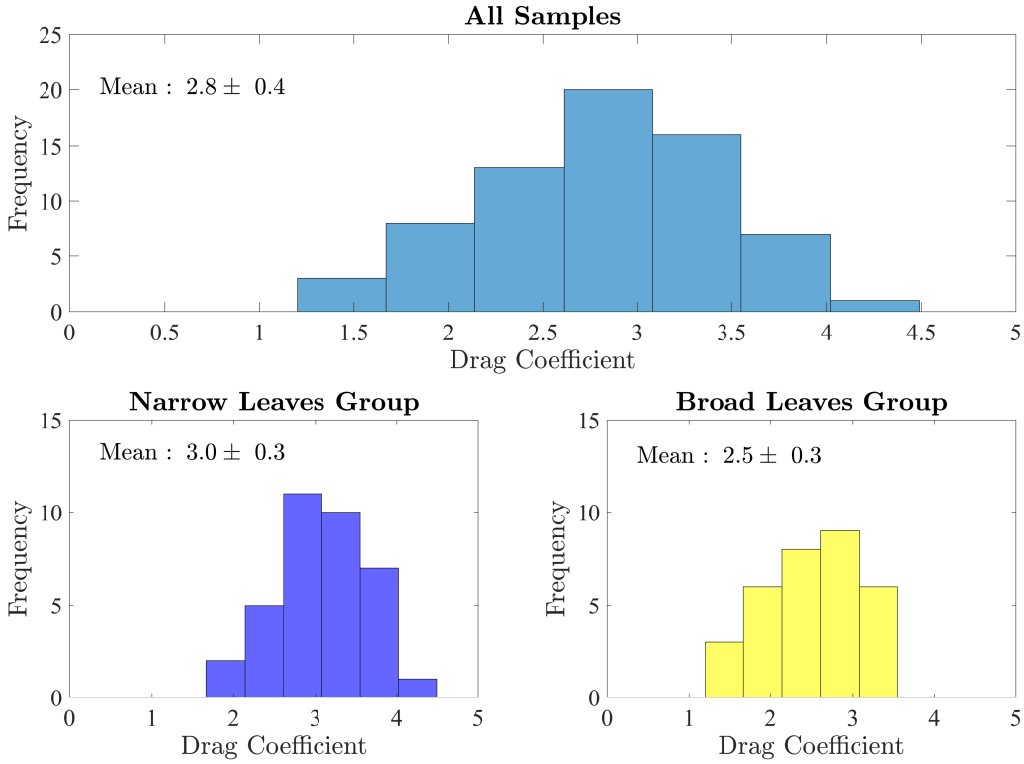


Fig. 8. Distribution of drag coefficients for all samples (top), samples with narrow leaves (bottom left), samples with broad leaves (bottom right).

5 Conclusion

This report documents a series of experiments implemented to determine the absorption coefficient, pressure loss, and the solid fraction of different types of vegetation sample configurations. The primary objective of this work was to calculate the drag coefficients of bulk vegetation that can be incorporated into CFD models. In addition to establishing drag coefficients of bulk vegetation, notable findings regarding vegetation structure and similarities between drag coefficients of plant species were also discovered from this work. It cannot be concluded, however, that the findings from this work applies to all bulk vegetation, but exclusively to the samples studied in these experiments.

To summarize, the findings of this work are as follows:

1. The calculated absorption coefficient for each sample demonstrated a strong relationship with its corresponding solid fraction.
2. The overall average drag coefficient of the bulk vegetation was found to be 2.8 with an expanded uncertainty of 0.4. The differences between the average drag coefficients of different plant species as well as the leaf type groups were shown to be significant, while still falling within the overall mean's uncertainty bound, suggesting that the overall average drag coefficient could be used as a constant value in CFD models of various plant types.

Falkenstein-Smith, Ryan; McGrattan, Kevin.
"Measurement of Drag Coefficients through Vegetation Canopy."
Paper presented at 15th International Conference and Exhibition on Fire Science and Engineering (Interflam 2019), London, United Kingdom.
July 1, 2019 - July 3, 2019.

Table 2. Drag coefficient summary of vegetation samples

Sample	β (%)	Position	C_d	Uncertainty	Sample	β (%)	Position	C_d	Uncertainty
Blue Spruce	1.9	A	3.6	0.5	Cypress	1.7	A	3.0	0.5
		B	3.1	0.5			B	3.2	0.5
		C	3.1	0.4			C	3.4	0.5
		D	3.0	0.4			D	3.0	0.4
	1.8	A	3.8	0.5		1.4	A	3.3	0.4
		B	3.0	0.4			B	2.9	0.4
		C	3.1	0.4			C	3.3	0.4
		D	3.6	0.4			D	3.8	0.4
	1.2	A	3.2	0.4		1.2	A	2.1	0.5
		B	2.6	0.3			B	3.2	0.3
		C	2.5	0.3			C	3.1	0.4
		D	2.8	0.3			D	3.3	0.4
	0.7	A	2.5	0.4		0.9	A	2.9	0.4
		B	2.3	0.4			B	3.9	0.4
		C	2.2	0.4			C	3.0	0.5
		D	2.2	0.4			D	3.8	0.5
White Pine	2.8	A	4.4	0.7	Fern	0.4	A	3.4	0.5
		B	2.8	0.4			B	3.1	0.4
		C	2.1	0.4			C	3.3	0.5
		D	3.6	0.5			D	2.6	0.4
Distylium	0.5	A	3.2	0.4	Red Holly	2.7	A	3.1	0.5
		B	2.9	0.4			B	2.6	0.4
		C	3.1	0.4			C	2.5	0.4
		D	3.4	0.4			D	2.7	0.4
	0.4	A	2.7	0.3		2.3	A	2.8	0.4
		B	2.3	0.3			B	2.0	0.3
		C	3.1	0.4			C	2.5	0.3
		D	2.9	0.4			D	1.8	0.3
	0.3	A	2.0	0.3		2.2	A	3.4	0.3
		B	1.5	0.2			B	2.2	0.2
		C	1.7	0.3			C	2.7	0.3
		D	2.6	0.3			D	2.5	0.3
						2.1	A	2.1	0.3
							B	1.6	0.3
							C	1.6	0.3
							D	1.7	0.3

Acknowledgments

The authors would like to thank Matthew Bundy and Artur Chernovksy of the National Fire Research Laboratory at NIST, who assisted in conducting these experiments and in processing the data.

References

[1] K. McGrattan, S. Hostikka, R. McDermott, J. Floyd, C. Weinschenk, and K. Overholt. *Fire Dynamics Simulator, Technical Reference Guide*. National Institute of Standards and Technology, Gaithersburg, Maryland, USA, and VTT Technical Research Centre of Finland, Espoo, Finland, sixth edition, September 2013. Vol. 1: Mathematical Model; Vol. 2: Verification Guide; Vol. 3: Validation Guide; Vol. 4: Software Quality Assurance.

[2] J. Cao, Y. Tamura, and A. Yoshida. Wind tunnel study on aerodynamic characteristics of shrubby specimens of three tree species. *Urban Forestry and Urban Greening*, 11(4):465–476, 2012.

[3] J. Jalonen and J. Järvelä. Estimation of drag forces caused by natural woody vegetation of different scales. *J. Hydrodynamics*, 26(4):608–623, 2014.

[4] G.J. Mayhead. Some Drag Coefficients for British Forest Trees Derived from Wind Tunnel Studies. *Agricultural Meterology*, 12:123–130, 1973.

[5] J. A. Gillies. Drag coefficient and plant form response to wind speed in three plant species: Burning Bush (Euonymus alatus), Colorado Blue Spruce (Picea pungens glauca.), and Fountain Grass (Pennisetum setaceum). *J. Geophysical Research*, 107(D24):ACL 10–1–ACL 10–15, 2002.

[6] H. Ishikawa, A. Suguru Amano, and Y. Kenta. Flow around a Living Tree. *JSME International Journal, Series B: Fluids and Thermal Engineering*, 49(4):1064–1069, 2006.

[7] F. Pimont, J.L. Dupuy, R. R. Linn, and S. Dupont. Validation of FIRETEC wind-flows over a canopy and a fuel-break. *Int. J. Wildland Fire*, 18(7):775–790, Oct. 2009.

[8] S. Dupont and Y. Brunet. Edge flow and canopy structure: a large-eddy simulation study. *Boundary Layer Meteorology*, 126(1):51–71, Jan. 2008.

[9] E. Mueller, W. Mell, and A. Simeoni. Large eddy simulation of forest canopy flow for wildland fire modeling. *Canadian J. Forest Res.*, 44(12):1534–1544, Jul. 2014.

[10] Emerson Process Management Rosemount Measurement, Chanhassen, Minnesota, USA. *Rosemount Annubar® Primary Flow Element Flow Test Data Book*, July 2009.

[11] R. Falkenstein-Smith, K. McGrattan, B. Toman, and M. Fernandez. Measurement of the Flow Resistance of Vegetation. Technical report, National Institute of Standards and Technology, April 2019.

[12] D. Lane and N. Salkin. Tukey's honestly significant difference (hsd). *N. Salkind der.), Encyclopedia of Research Design içinde, Thousand Oaks, CA: SAGE Publications*, 1:1566–1571, 2010.

[13] B. Toman and A. Possolo. Laboratory effects models for interlaboratory comparisons. *Accreditation and Quality Assurance*, 14(10):553–563, October 2009.

[14] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis, 2nd Edition*. Chapman and Hall/CRC, 2013.

[15] David Lunn, David Spiegelhalter, Andrew Thomas, and Nicky Best. The bugs project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067, October 2009.

The Chemical Structure of Medium-Scale Pool Fires

Ryan Falkenstein-Smith, Kunhyuk Sung, Jian Chen, and Anthony Hamins,
National Institute of Standards and Technology, Gaithersburg, MD 20899-8664, USA

1. Introduction

Use of fire modeling, such as the Fire Dynamic Simulator[1], in fire protection engineering has increased dramatically during the last decade due to the development of practical computational fluid dynamics fire models and the decreased cost of computational power. Today, fire protection engineers use models to design safer buildings, nuclear power plants, aircraft cabins, trains, and marine vessels to name a few types of applications. To be reliable, the models require validation, which involves an extensive collection of experimental measurements. An objective of this report is to provide data for use in fire model evaluation by the research community.

A pool fire is a fundamental combustion configuration of interest in model development. In pool fires, the fuel surface is isothermal, flat and horizontal, which provides a well-defined and straightforward setup for testing models and furthering the understanding of fire phenomena. In moderate and large-scale pool fires, radiative heat transfer is the dominant mechanism of heat feedback to the fuel surface. Species concentrations and temperatures have a significant influence on the radiative heat transfer. A zone of particular interest is the fuel rich-core between the flame and the pool surface, where gas species can absorb energy that would otherwise have been transferred to the fuel surface. Few studies in the literature studies have reported local chemical species measurements, which provide a deep understanding of the chemical structure of a pool fire and provide insight on critical kinetic, heat, and mass transfer processes.

The purpose of this study is to characterize the spatial distribution of stable gas-phase chemical species in a moderate-scale liquid pool fire steadily burning in a well-ventilated quiescent environment. Here, methanol is selected as the fuel. Fires established using methanol are unusual as no carbonaceous soot is present or emitted. This creates a particularly useful testbed for fire models and radiation sub-models that consider emission by gaseous species - without the confounding effects of blackbody radiation from soot.

In this study, measurements are made in a 0.30 m diameter methanol pool fire. This particular fire was selected for study since the measurements complement the results from previous studies, including analyses of the mass burning rate, the temperature and velocity fields, radiative emission, flame height, and pulsation frequency[2-3]. Additional characterization of this fire enables a more comprehensive understanding of its detailed structure, enhancing the understanding of fire physics.

2. Experimental Method

The experimental setup used in this work has been documented previously[3-6]. Experiments were conducted under a canopy hood surrounded by a 2.5 m x 2.5 m x 2.5 m enclosure made of a double-mesh screen wall. The walls of the enclosure were formed by a double layer of wire-mesh screen (5 mesh/cm) that reduced the influence of compromising room flows that could disrupt the pool fire's flow field. All measurements were made once the burning conditions, specifically the mass burning rate, reached steady-state, achieved approximately 10 min after ignition.

Falkenstein-Smith, Ryan; Sung, Kunhyuk; Chen, Jian; Hamins, Anthony.
"The Chemical Structure of Medium-Scale Pool Fires."
Paper presented at 15th International Conference and Exhibition on Fire Science and Engineering (Interflam 2019), London, United Kingdom.
July 1, 2019 - July 3, 2019.

2.1 Pool Burner

A circular, stainless-steel pan with an outer diameter of 0.30 m, a depth of 0.15 m, and a wall thickness of 0.0016 m was used as the pool burner. The burner was placed within an overflow basin, which extended 3 cm beyond the burner wall. The burner was fitted with legs such that the burner rim was positioned 0.3 m above the ground. The bottom of the burner was maintained at a constant temperature by flowing water (20 °C ± 3 °C) through a 3 cm section on the bottom of the fuel pan.

While burning, the fuel level was monitored by a camera with a zoom lens to allow visual observations, displayed on a 50 cm monitor, of the barely discernable dimple made from the fuel-level indicator on the fuel surface. The fuel level indicator was positioned near the center of the burner as shown in Figure 1. To preserve a consistent mass burning rate, the fuel level was maintained 10 mm below the burner rim, following conditions used in previous studies. Fuel to the burner was gravity fed from a reservoir positioned on a mass load cell located outside the enclosure and monitored by a data acquisition system. The fuel flow was manually adjusted using a needle valve. The expanded uncertainty with a coverage factor of 2 of the fuel level was estimated to be 0.5 mm for these experiments.

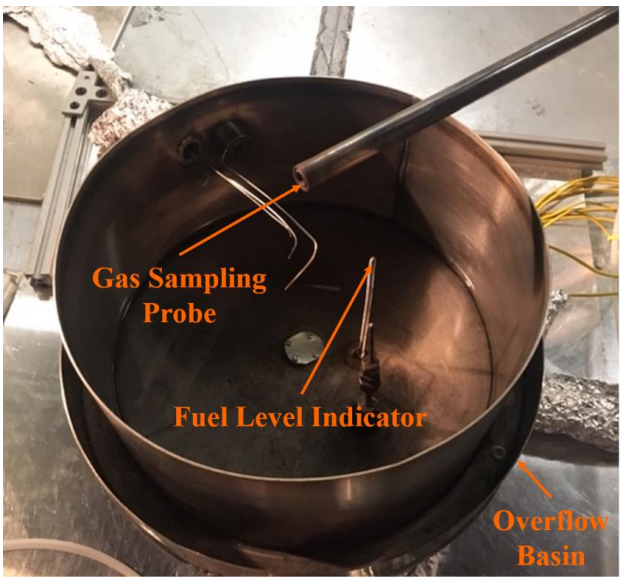


Figure 1. Pool Burner, 30 cm in diameter, with fuel level indicator, overflow section, and quenching probe

2.2 Gas Species Measurement

Gas-species measurements were made using an Agilent 5977E Series Gas Chromatograph/Mass Spectrometer System (GC/MS) fitted with a thermal conductivity detector (TCD).* The GC/MS was able to quantify a variety of stable reactants, intermediates, and combustion product species collected from the pool fire. The GC was conjured with a 2 ml sampling loop. Chromatographic separation of species was achieved using a Select for Permanent Gases-Dual Column (CP7430) comprised of mole-sieve and Porapak Q columns working in parallel and using a helium carrier gas. The sample analysis time was 33 min wherein the carrier gas flow leading into the TCD and MS was 3.0 ml/min and 1 ml/min, respectively. During the analysis, the GC oven temperature was maintained at 30°C for 10 min, then ramped at 8°C/min for 20 min until a temperature of 190°C was obtained.

* Certain commercial products are identified in this report in order to specify adequately the equipment used. Such identification does not imply recommendation by the National Institute of Standards and Technology, nor does it imply that this equipment is the best available for the purpose.

Figure 2 displays the flow diagram for gas sampling into the GC/MS. After achieving steady-state burning conditions, a vacuum pump, located downstream from the GC/MS, was initiated. Gas samples were collected using a quenching probe. The quenching probe was composed of two concentric, stainless-steel tubes with outer annular coolant flow and inner, extracted, gas-sample flow. The inner and outer tube diameters were 7.9 mm and 16 mm, respectively. Water maintained at 90°C was used to flow through the sampling probe for the duration of the experiment. The remainder of the sampling line leading into the GC was heated with electrical, heating tape to prevent condensation of water and methanol through the line.

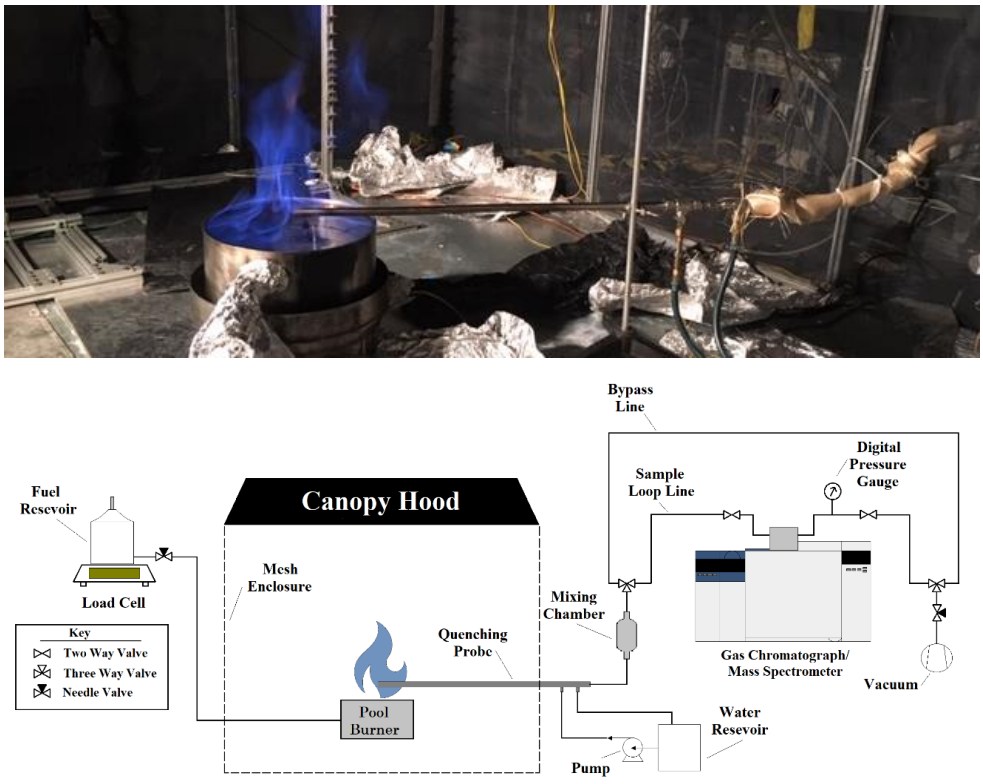


Figure 2. Photo of the experiment with the sampling probe in the fire (top) and schematic drawing of the flow diagram (bottom) for extractive sampling and transport of gas samples from the pool fire to the GC/MS

The sample line consisted of a 150 ml mixing chamber positioned upstream from the GC/MS. To reduce the time required to fill the mixing chamber, the sample line was split downstream into a 6.40 mm diameter "bypass" and 1.7 mm diameter "sample loop" line. For the first four minutes of sampling, gas was collected through the "bypass" line at a rate of 3.0 L/min. After the sampling path was switched to the "sample loop" line, the flow was reduced to 0.75 L/min for six minutes leading up to injection. When the sampling period was completed, the two-way valve located downstream of the sample loop line was closed. The gas sample was held within the sample loop line for one minute before injection to allow the pressure to equilibrate. Pressure was monitored using a digital pressure gauge located downstream of the GC/MS.

TCD and MS gas calibrations were conducted using commercial gas phase calibration standards. Water and Methanol calibrations were conducted using the bubbler setup shown in Fig. 3. Nitrogen, acting as a carrier gas, flowed into a liquid bath and then into the sample loop of the GC/MS. The nitrogen carrier gas flow for all calibrations was estimated to be 30 ml/min ± 3.0 ml/min. A thermocouple was placed

at the liquid surface to capture the bath temperatures needed to determine vapor pressure. Liquid bath temperatures were controlled using a heating plate positioned underneath the insulated bubbler. Liquid calibrations were conducted once the bath temperature achieved steady state (approximately 1 hour).

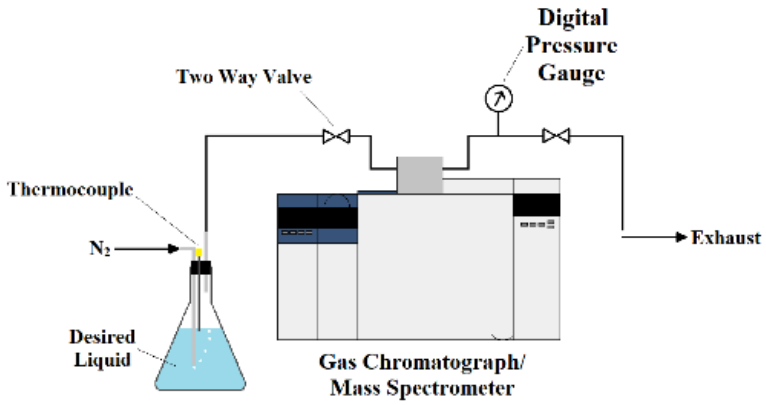


Figure 3. Flow diagram for bubble calibration system used for water and methanol

All measurements were repeated at least twice at each location along the centerline of the pool fire. Gas species concentration measurements made at the same location were averaged. The variance in the gas species volume fraction varied between location and species. The combined uncertainty was calculated from the standard deviation of the repeated measurements multiplied by a coverage factor of two, representing a 95 % confidence interval. A detailed description of the uncertainty analysis for the gas species measurement is described in detail elsewhere[7].

3. Results and Discussion

Figure 4 shows two photos of the pulsing methanol pool fire. The fire shape fluctuated during its pulsing cycle with uniformly curved flame sheets present at the burner rim that rolled towards the fire centerline to form a long and narrow plume. The observed dynamic fire shape is consistent with previous descriptions[3,8-9].

Figure 4. Images of pulsing methanol pool fire (30 cm diameter)

3.1 Gas Species Concentrations

Table 1 and Figure 5 display the averaged volume fraction, $\bar{X}_i$, of the significant species concentrations made at various heights along the fire's centerline and relative to the fuel surface. The species measured included the reactants, Methanol (CH_3OH) and oxygen (O_2), combustion products such as water (H_2O) and carbon dioxide (CO_2), combustion intermediates such as carbon monoxide (CO), hydrogen (H_2), methane (CH_4), and inert gases such as Nitrogen (N_2) and Argon (Ar). As expected, the fuel volume fractions were highest, and the oxygen volume fraction was the lowest close to the fuel surface. The product species were found to have a maximum volume fraction at approximately 4 cm above the fuel surface. The intermediate gas species were found to have peaked at approximately 2 cm above the fuel surface. Inert gases were shown to increase with the distance from the fuel surface. It was also found that the gas sampled from the centerline nearly mimicked the composition of air at approximately 60 cm above the fuel surface.

Table 1. Concentrations of major species as a function of height relative to the pool surface

Position	$\bar{X}_{CH_3OH}$	$\bar{X}_{O_2}$	$\bar{X}_{CO_2}$	$\bar{X}_{H_2O}$	$\bar{X}_{CO}$	$\bar{X}_{H_2}$	$\bar{X}_{N_2}$	$\bar{X}_{Ar}$
1.0 cm	24 ± 4.8	2.2 ± 0.2	4.3 ± 0.4	15 ± 7.6	5.9 ± 0.6	5.3 ± 0.6	42 ± 4.4	0.4 ± 0.1
2.0 cm	20 ± 3.9	2.1 ± 0.2	4.6 ± 0.5	17 ± 8.5	6.6 ± 0.7	5.6 ± 0.7	44 ± 4.8	0.6 ± 0.1
4.0 cm	11 ± 1.9	3.4 ± 0.3	5.9 ± 0.6	16 ± 8.0	6.1 ± 0.6	4.5 ± 0.5	52 ± 5.3	0.6 ± 0.1
6.0 cm	9.4 ± 1.7	5.1 ± 0.5	5.7 ± 0.5	14 ± 6.8	4.2 ± 0.4	3.1 ± 0.3	57 ± 5.3	0.7 ± 0.1
14 cm	5.1 ± 1.9	8.8 ± 0.8	5.0 ± 0.4	12 ± 5.2	1.7 ± 0.2	1.0 ± 0.1	65 ± 5.5	0.8 ± 0.1
20 cm	*	13 ± 1.0	3.6 ± 0.3	9.4 ± 3.3	0.3 ± 0.0	0.1 ± 0.0	72 ± 5.4	0.8 ± 0.1
30 cm	*	17 ± 1.2	2.1 ± 0.1	4.3 ± 1.1	Trace	*	76 ± 5.3	0.9 ± 0.1
60 cm	*	19 ± 1.4	0.9 ± 0.1	*	Trace	*	79 ± 5.5	0.9 ± 0.1

* Species not detected; Trace indicates concentrations less than 1000 ppm

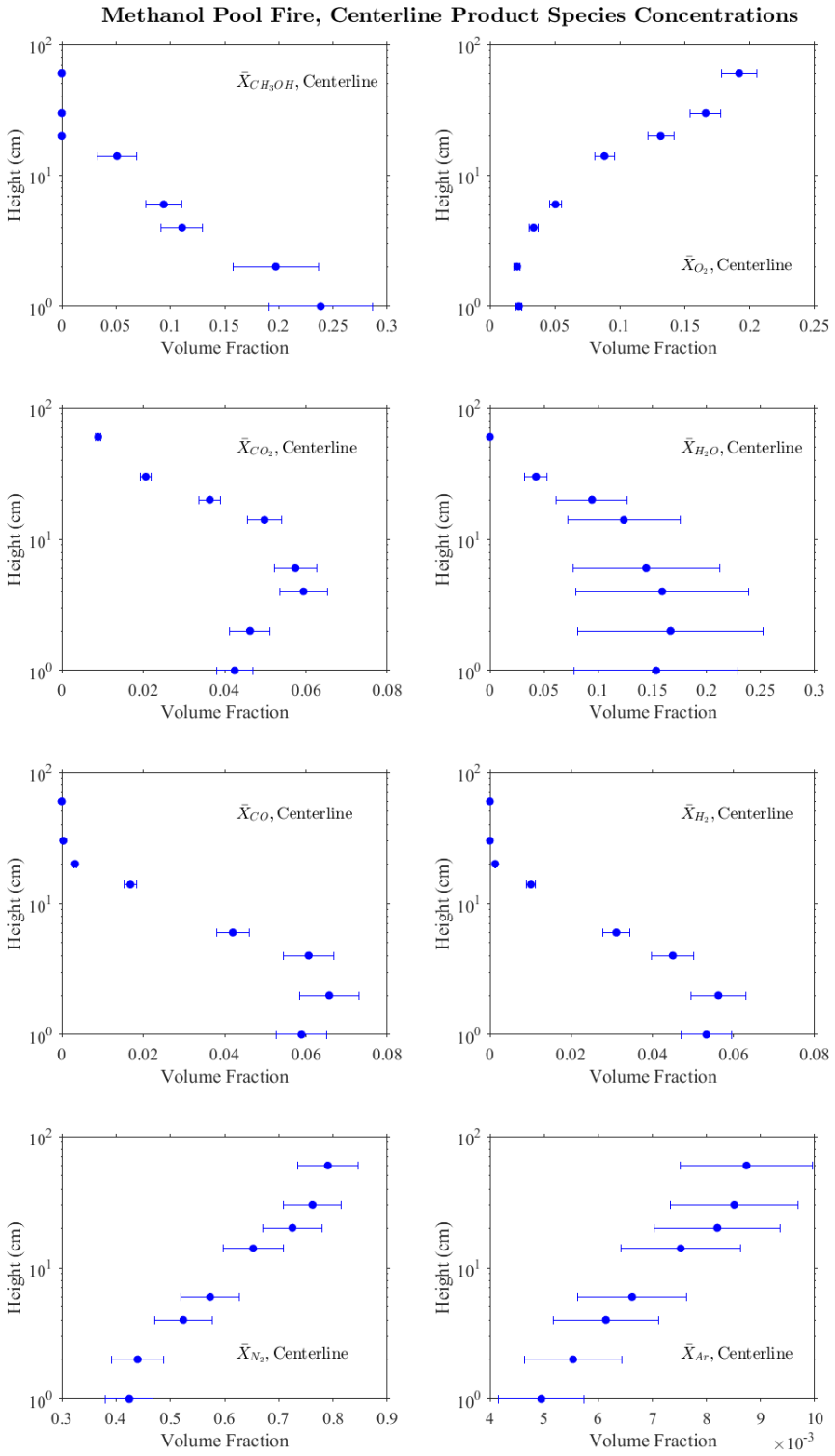


Figure 5. Volume fraction of major gas species as a function of height relative to the pool surface with uncertainty defined as the standard deviation of the repeated measurements with a coverage factor of 2

4. Conclusion

In summary, time-averaged local measurements of gas species concentrations were made to characterize the structure of a 30 cm diameter, methanol, pool fire steadily burning in a quiescent environment. These local measurements are essential to providing insight into the complex dynamics and chemical structure of medium-scale pool fires. It is anticipated that the data will be useful for model validation.

5. Acknowledgments

The authors would like to acknowledge Kimberly Harris of the Gas Sensing Metrology Group at NIST, who assisted in the developing the species calibration method used in the study.

References

[1] K. McGrattan, S. Hostikka, R. McDermott, J. Floyd, C. Weinschenk, and K. Overholt. *Fire Dynamics Simulator, Technical Reference Guide*. National Institute of Standards and Technology, Gaithersburg, Maryland, USA, and VTT Technical Research Centre of Finland, Espoo, Finland, sixth edition, September 2013. Vol. 1: Mathematical Model; Vol. 2: Verification Guide; Vol. 3: Validation Guide; Vol. 4: Software Quality Assurance.

[2] S.J. Fischer, B. Hardouin-Duparc, and W.L. Grosshandler. The structure and radiation of an ethanol pool fire. *Combustion and flame*, 70(3): 291–306, 1987.

[3]A. Hamins and A. Lock. *The Structure of a Moderate-Scale Methanol Pool Fire*. US Department of Commerce, National Institute of Standards and Technology, 2016.

[4] A. Hamins, M. Klassen, J. Gore, S. Fischer, and T. Kashiwagi. Heat Feedback to the Fuel Surface in Pool Fires, Combust. Sci. Tech., 97, 37-62 (1993).

[5] A. Hamins, M. Klassen, J. Gore, T. and Kashiwagi, *Estimate of Flame Radiance via a Single Location Measurement in Liquid Pool Fires*, Combustion and Flame, 86:223-228 (1991).

[6] A. Hamins, T. Kashiwagi, and R. Buch, *Characteristics of Pool Fire Burning, in Fire Resistance of Industrial Fluids*, ASTM STP 1284 (Eds: G. Totten and J. Reichel), American Society for Testing and Materials (ASTM) Publication Number 04-012840-12, W. Conshocken, PA, pp. 15-41 (1995).

[7] A.Lock, M. Bundy, E.L. Johnsson, A. Hamins, G.K. Ko, C. Hwang, P. Fuss, and R. Harris, *Experimental Study of the Effects of Fuel Type, Fuel Distribution, and Vent Size on Full-Scale Underventilated Compartment Fires in an ISO 9705 Room*, NIST Technical Note 1603, National Institute of Standards and Technology, Gaithersburg, MD, October 2008.

[8] E.J. Weckman and A. Sobiesiak, *The Oscillatory Behavior of Medium-Scale Pool Fires, Proceedings of the Twenty-Second Sym. (Int.) on Combustion*, The Combustion Institute, 1299-1310 (1988).

[9] H. R. Baum, and B. J. McCaffrey, *Fire Induced Flow Field - Theory and Experiment, Proceedings of the Second International Symposium on Fire Safety Science*, Hemisphere, New York, 129-148 (1989).

Design-for-Cost – An Approach for Distributed Manufacturing Cost Estimation

Minchul Lee[1] and Boonserm (Serm) Kulvatunyou[1]

[1]Systems Integration Division, National Institute of Standards and Technologies
{ minchul.lee, serm }@nist.gov

Abstract. Researches have shown that design changes cost more in later stages of product development. Therefore, companies adopt Design-for-X methods to optimize product designs from many aspects in the early design stage. Despite such efforts, they often encounter several design changes during the commission of the production; and one of the main reasons is a failure to meet their target cost. Accurately estimating cost in the early design stage is difficult due to insufficient information. In particular, as production becomes more distributed cost estimation is also more difficult because information is more distributed. This paper introduces a cost estimation method to address this problem. It describes a distributed manufacturing situation and a cost breakdown framework. A use case is provided to illustrate how the framework allows for supply-chain cost negotiation and design adjustments in the early design stage.

Keywords: Cost Estimation, New Product Introduction, Cost Breakdown Approach, Design-for-Cost, Supply Chain Management.

1 Introduction

Researches have shown that design changes cost more in later stages of product development [1]. Therefore, companies adopt Design-for-X methods to optimize the product design from many aspects such as quality, time to delivery and cost in the early design stage. Companies have strong interest in the ability to accurately estimate manufacturing cost earlier in the design stage, because executives pay a lot of attention to profit.

In the New Product Introduction (NPI) process, companies typically set target market, volume, price, and manufacturing cost along with the design and functions of a product at the first stage [2]. They are very interested in maintaining the profit margin; therefore, target manufacturing costs are validated at every NPI stage. Despite this general practice, unexpected cost usually shows up at the commission of production because the current cost estimation approach is insufficient for today's distributed manufacturing environment.

Traditionally, manufacturers[1] have most of the information necessary for cost estimation. It is not the case for distributed manufacturing. Manufacturers need to interact

[1] In this paper, *manufacturers* refer to the organizations that design and/or produce products or components that need subassemblies from *supplier* organizations. The two roles are played

Lee, Minchul; Kulvatunyou, Boonserm.
"Design-for-Cost – An Approach for Distributed Manufacturing Cost Estimation."
Paper presented at APMS 2019 Conference, Advances in Production Management Systems, Austin, TX, United States. September 1, 2019 - September 5, 2019.

with suppliers, estimate component costs, and consider delivery and packaging costs from the early design stage. Cost estimation approach for distributed manufacturing needs to enable manufacturers to negotiate with suppliers and come up with supply chain strategies to reduce cost. For example, in addition to the typical design adjustments to reduce material and manufacturing process costs, manufacturers may find suppliers who can use the same material and purchase the material on behalf of all the suppliers to receive larger bulk-buying discount. Without proper cost break down in the estimation, manufacturers will have difficulty negotiating with suppliers as they do not know which cost elements can be reduced.

This paper proposes a framework for manufacturing cost estimation in the early design stage for the case of distributed manufacturing. The framework reduces risks in encountering unexpected cost at the commission of production that could result in a costly design change. The rest of paper is organized as follows. First, a literature review on manufacturing cost estimation is given. Then, our cost estimation framework is outlined. Finally, a case study showing cost estimation of a supplied component is illustrated followed by a conclusion and future work.

2 Literature Review

In this chapter, a summary of existing manufacturing cost estimation methods and cost elements is provided.

2.1 Cost Estimation Methods

Manufacturing cost estimation methods can generally be divided into two groups: qualitative methods and quantitative methods [3].

Qualitative methods are based on a comparative analysis between a new product and similar products manufactured previously. On the other hand, quantitative methods are based on a detailed cost analysis of product design, its features, and corresponding manufacturing processes instead of simply relying on the past data or tribal knowledge of an estimator. The qualitative methods do not provide a cost break down that can be used to understand cost elements to the benefit of design change and cost negotiation.

There are several types of quantitative method including Operation-based approach [4], Feature-based approach [5] and Break-down approach [6]. Operation-based approach mainly estimates cost in terms of types of operations and considers material cost, factory expenses and manufacturing processing cost as part of the costs associated with time of performing operations. This approach focuses on accurate estimation of the manufacturing processing cost but provides less detailed consideration on other cost elements.

by organizations in a distributed manufacturing chain. For example, GM (a manufacturer) designs and produces cars which requires instrument panel assembly from Delphi (a supplier). Delphi on the other hand can be a manufacturer ordering electrical harness from a supplier.

Lee, Minchul; Kulvatunyou, Boonserm.
"Design-for-Cost – An Approach for Distributed Manufacturing Cost Estimation."
Paper presented at APMS 2019 Conference, Advances in Production Management Systems, Austin, TX, United States. September 1, 2019 - September 5, 2019.

The feature-based cost estimation approach identifies cost-related features of the design and respectively estimates their costs. However, existing approaches in this category only considered conventional machining process.

The break-down approach breaks down manufacturing cost into cost elements. An estimation is applied to each cost elements. The estimated manufacturing cost is a sum of all cost elements incurred during the production cycle. Cost elements include material cost, manufacturing process cost, maintenance cost, and repair cost. Additionally, [6] added insurance cost, [7] added overhead costs, and [8] calculated the manufacturing process cost based on the hourly usage of machinery.

For more accurate cost estimation, we adopt a cost break-down approach. However, our research focuses on estimating cost for distributed manufacturing. For that, we have to look into what data is available for the manufacturer; and it is necessary to extend the scope of cost elements such as packaging and delivery cost. The scope of this paper and associated cost elements is described in 2.2.

2.2 Cost Elements

According to [9], selling price consisted of material, labor, indirect, selling and administrative expense and profit as shown in **Fig. 1**.

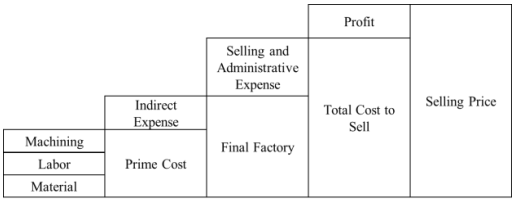


Fig. 1. Elements of Cost

Most researches focus on *Final Factory Cost*, however, it is necessary to widen the scope to *Selling Price* to include packaging and delivery cost in the *Total Cost to Sell* in order to estimate cost of a distributed manufacturer. To enable detailed cost analysis, cost elements for machining, tool, and defects should be added to *Prime Cost*. In conclusion, cost elements for distributed manufacturing are shown in **Fig. 2**.

Material	Tool	Machining	Labor	Defect	Packaging	Delivery	Selling & Administrative	Profit
Selling Price = Distributed Manufacturing Cost								

Fig. 2. Cost Elements for Distributed Manufacturing

3 Cost Estimation Framework

In this chapter, we introduce a cost estimation framework for a supplied component. We suppose that design of the component starts after target market, volume, and price are set as they are needed for the cost estimation.

Lee, Minchul; Kulvatunyou, Boonserm.
"Design-for-Cost – An Approach for Distributed Manufacturing Cost Estimation."
Paper presented at APMS 2019 Conference, Advances in Production Management Systems, Austin, TX, United States. September 1, 2019 - September 5, 2019.

Several researches provided logic (cost elements and formula) for estimating costs. However, they can be difficult to apply in the early design stage of distributed manufacturing because data are not available. Therefore, it is necessary to incrementally increase the accuracy of cost estimation as shown in **Fig. 3** - starting from using logic requiring least information and add more details and data as they become available. The approaches to update logic and improve the accuracy include further breaking cost elements, decompose parameters in formula, and collecting more data and update the database. Ways to collect data include investigating trends in the market, prototyping the component internally, contacting the supplier, contacting the equipment makers, and contacting raw material suppliers.

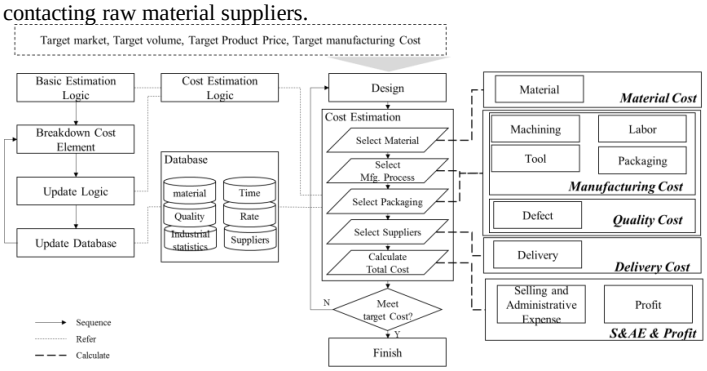


Fig. 3. Cost Estimation Framework for Distributed Manufacturing

As shown in **Fig. 2**, Selling Price of supplier (or supplied component) needs to be estimated. Therefore, the cost of the designed part is estimated considering the material, the manufacturing process, the packaging method, and the anticipated supplier, and then compare with the target cost. Right hand side of **Fig. 3** shows which cost elements (in **Fig. 2**) are related to cost estimation steps.

In the early stages of development, since every aspect of the design is not decided, such as the raw material, manufacturing method, assumptions have to be made on the parameters of cost elements. Therefore, only as the design of the product becomes more concrete, the cost can be estimated more accurately.

4 Cost Estimation Method and Cost Breakdown

In this chapter, we introduce basic formula for each cost element and discuss their essential aspects. The formula may need to be adjusted for different supplied components and situations. Each of the cost elements is estimated per a part.

Lee, Minchul; Kulvatunyou, Boonserm.
"Design-for-Cost – An Approach for Distributed Manufacturing Cost Estimation."
Paper presented at APMS 2019 Conference, Advances in Production Management Systems, Austin, TX, United States. September 1, 2019 - September 5, 2019.

4.1 Material Cost

Most of the researches estimate material cost with unit cost and amount and a basic estimation logic of material cost C_{mat} is given by

$$Material\ Cost\ C_{mat} = m_t \times U_{mat} \tag{1}$$

where m_t is the total amount of the raw material and U_{mat} is the unit cost for raw material However, the amount of material should be broken down into the net material, m_n, and loss material, m_l, to present an opportunity to reduce the loss material. The design engineer may improve the design or the manufacturer may work with supplier to improve the process to reduce the loss material. Loss materials are, for example, for preheating in injection molding process and for chips or scraps in NC machining. Thus, the material cost C_{mat} which characterizes material loss is given by:

$$Material\ Cost\ C_{mat} = (m_n + m_l) \times U_{mat} \tag{2}$$

4.2 Machining, Labor and Tool Costs

Generally, the basic estimation formula for machining is expressed as the product of machine rate and machining time [6].

$$Machining\ Cost\ C_{mcn} = r_m \times t_m \tag{3}$$

where r_m is the machine rate and t_m is the machining time. Machine rate is usually calculated by dividing the cost to operate a machine or machines for the needed processing duration by the duration itself. Details of machine rate estimation is beyond the scope of this paper, but it can be calculated approximately with depreciation, electricity and maintenance cost. Machining time can be subdivided into setup time (t_s) operation time(t_o) and nonoperation (idle time and down time) time (t_{no}) [10].

$$Machining\ Cost\ C_{mcn} = r_m \times (t_s + t_{o\,+}\, t_{no}) \tag{4}$$

Labor cost which is similar with machining cost is given by

$$Labor\ Cost\ C_l = r_l \times t_l \tag{5}$$

where r_l is the labor rate and t_l is the labor operation time. A labor rate is the cost of labor that is used to deriving the costs of various activities or products directly related to manufacturing. Calculating accurate labor cost can be difficult for distributed manufacturing because every supplier has different wage and compensation. Average labor rate of the industry can be used and is a good reference point for negotiation. The average labor operation time can be obtained by dividing the total number of products made during the day by the work time per day. However, it is also possible to apply the technique of analyzing the operation time such as Modular Arrangement of Predetermined Time Standards (MODAPTS) [11].

Tool cost is the cost of devices such as a mold or a jig, which is calculated by dividing the price of the device by the target production quantity.

Lee, Minchul; Kulvatunyou, Boonserm.
"Design-for-Cost – An Approach for Distributed Manufacturing Cost Estimation."
Paper presented at APMS 2019 Conference, Advances in Production Management Systems, Austin, TX, United States. September 1, 2019 - September 5, 2019.

$$Tool\ Cost\ C_{tool} = \frac{Tools\ Prices}{Target\ production\ volume} \quad (6)$$

4.3 Defect Cost

As every manufacturing process has failures, defect cost needs to be considered. Defect rate is difficult to predict before start of manufacturing and it can vary between suppliers. Therefore, the defect rate is typically set to the same as or lower than an average of historical defect rates. In order to predict the defect cost, defect cost C_{dft} is given by

$$Defect\ Cost\ C_{dft} = (C_{mat} + C_{mcn} + C_l) \times r_d \quad (7)$$

where r_d is the defect rate.

4.4 Packaging Cost

Packaging cost includes all the materials and processes needed for delivering the supplied component to the manufacturer.

$$Packaging\ Cost\ C_{pkg} = p_b \div n_p + r_l \times t_p \quad (8)$$

where p_b is the cost for package box, n_p is the number of parts for a box and t_p is the packing time.

4.5 Delivery Cost

Delivery cost is the cost of delivering the packaged product from the supplier to the manufacturer. The method of estimating cost may vary depending on the transportation, but the formula below includes essential parameters for cost reduction analysis.

$$Delivery\ Cost\ C_{dlv} = r_d \times d \div n_t \div n_p \quad (9)$$

where r_d is the delivery rate, d is the distance from supplier to buyer and n_t is the number of boxes in a transport.

4.6 Selling and Administrative Cost and Profit

These cost elements can be estimated as percentage over other cost elements. Selling and administrative (S&A) cost also includes research and development. These ratios vary by industries and suppliers and may be obtained from industry statistics such as in [12-13]. The S&A ratio can be higher for a supplier with an R&D center, such as a PCB manufacturer. In addition, the profit ratio can be negotiable, and it may increase for favorable partners such as those delivering consistent quality components and on-time delivery. The basic formulas for the two cost elements are:

$$\text{Selling and Administrative Cost } C_{sac} = (C_{mat} + C_{mcn} + C_l + C_{pkg} + C_{dlv}) \times r_{sac} \quad (10)$$

$$\text{Profit } C_{prf} = (C_{mat} + C_{mcn} + C_l + C_{pkg} + C_{dlv} + C_{sac}) \times r_{prf} \quad (11)$$

Lee, Minchul; Kulvatunyou, Boonserm.
"Design-for-Cost – An Approach for Distributed Manufacturing Cost Estimation."
Paper presented at APMS 2019 Conference, Advances in Production Management Systems, Austin, TX, United States. September 1, 2019 - September 5, 2019.

where r_{sac} is the selling and administrative ratio and r_{prf} is the profit ratio.

5 Use case

The use case is a scenario in which a refrigerator manufacturer needs an injection-molding parts from a supplier. After the supplier receives the drawings, molds are manufactured. And after the injection-molding part production and quality inspections are completed, they are put in protective tapes and delivered in boxes. Below, we discuss cost estimations and cost reductions experienced in this use case.

The component used 120g of abs resin; therefore, according to (1) is:

$$Material\ Cost\ C_{mat} = m_t\ (120g) \times U_{mat}\ (\$2.7/kg) = \$0.324$$

However, it was determined that if two parts were produced at the same time in a single mold, machining, labor, and tool costs were reduced despite an increase in material cost. The increase in material cost came from additional resin needed for the sprue and runner (12g) in the mold. Hence, material cost using (2) is:

$$Material\ Cost\ C_{mat} = m_t\ (120g + 12/2) \times U_{mat}\ (\$2.7/kg) = \$0.340$$

The price of the mold to produce the product is about $30,000 and can be used 200,000 times. However, since we produce two parts in one mold,

$$Tool\ Cost\ C_{tool} = \frac{\$300{,}000}{200{,}000} \div 2 = \$0.75$$

Machine rate for injection is calculated to be $24 per hour and the machining time is 40 seconds for two pieces.

$$Machining\ Cost\ C_{mcn} = r_m(\$24/hour) \times t_{mcn}\ (40s / 2) = \$0.13$$

The operator extracts the injected parts from the machine and performs visual inspection. The labor rate is $30 per hour and has the same tact time as machining time. Therefore, applying (3),

$$Labor\ Cost\ C_l = r_l(\$30/hour) \times t_l\ (40s / 2) = \$0.17$$

Since historically the average failure rate of the supplier in the use case was 2%,

$$Defect\ Cost\ C_{dft} = (C_{mat}(\$0.340) + C_{mcn}(\$0.33) + C_l(\$0.37)) \times r_d(2\%) = \$0.01$$

After the manufacturing is completed, tape is attached, and a total of 100 parts are inserted into a $2 container, which takes about 20 seconds per part. Therefore,

$$Packaging\ Cost\ C_{pkg} = p_b(2) \div n_p(100) + r_l(\$30/hour) \times t_p(20s) = \$0.187$$

A truck delivers parts with 400 boxes for 20km,

$$Delivery\ Cost\ C_{dlv} = r_d\ (\$20/km) \times d(20) \div n_t(400 \times 100) = \$0.01$$

Since the average S&A rate is 7% and profit rate is 8%, they cost $0.11 and $0.14.

And finally, the total cost is a summation of all the cost elements: $1.85 per part.

Discussion: If the estimated cost cannot meet the target cost, it is necessary to consider changing to cheaper raw materials, increasing the number of parts in the mold, or changing the packaging method to reduce the cost.

The estimated costs cannot be guaranteed to be available from the supplier but can be negotiated with supplier based on the estimation. For example, it may be possible to negotiate the defect rate, S&A rate, etc.; or to come up with a supply chain strategy such as bulk buying raw materials at a lower price on behalf of multiple suppliers, arrange logistics services, etc.

Lee, Minchul; Kulvatunyou, Boonserm.
"Design-for-Cost – An Approach for Distributed Manufacturing Cost Estimation."
Paper presented at APMS 2019 Conference, Advances in Production Management Systems, Austin, TX, United States. September 1, 2019 - September 5, 2019.

6 Conclusion

In this paper, we introduced a method and a use case to estimate cost for distribute manufacturing components. In order to estimate cost of distribute manufacturing, *selling price* should be considered instead of *final factory cost*. We also introduced cost elements of the selling price for a more accurate cost estimation. Through this, a framework is proposed that incrementally enhance the detail of cost estimation during the design process. A use case illustrated cost estimation of a component and discussed how parameters identified in cost elements may be used for adjusting design and negotiating with suppliers.

Future works include integrating the quantitative cost estimation approaches described in this paper with qualitative approaches. In this way machining time and working time can be better predicted based on data kept in a database of cost estimation. Therefore, defining data schema for such database is an important research topic.

References

1. Ulrich, K.T., D.J. Ellison, Product Design and Development, McGraw-Hill, New York (1999).
2. Riitta K., Gautam A.: Something Old, Something New: A Longitudinal Study of Search Behavior and New Product Introduction. Academy of Management Journal 45(6), 1183-1194 (2002).
3. Adnan N., Jian D.: Product Cost Estimation: Technique Classification and Methodology Review. Journal of Manufacturing Science and Engineering 128(2), 563-575 (2005).
4. Shehab, E. M., Abdalla, H. S.: Manufacturing Cost Modeling for Concurrent Product Development. Robotics and Computer-Integrated Manufacturing 17(4), 341-353 (2001).
5. Zhang, Y. F., Fuh, J. Y. H., and Chan, W. T.: Feature-Based Cost Estimation for Packaging Products Using Neural Networks. Computer Industry 32, 95–113(1996).
6. Son, Y. K.: A Cost Estimation Model for Advanced Manufacturing Systems," International journal of Production Research 29(3), 441–452(1991).
7. Bernet, N., Wakeman, M. D., Bourban, P. E., and Månson, J. A. E.: An Integrated Cost and Consolidation Model for Commingled Yarn Based Composites. Composites Part A: Applied Science and Manufacturing 4, 495–506(2002).
8. Ostwald, P. F., Engineering Cost Estimating, Prentice Hall, Englewood Cliffs, NJ (1992).
9. Colin D., Management and Cost Accounting, Cengage Learning EMEA (1985).
10. Jung, J.: Manufacturing Cost Estimation for Machined Parts Based on Manufacturing Features, Journal of Intelligent Manufacturing 13(4), 227–238(2002).
11. International MODAPTS Association, http://www.modapts.org/ , last accessed 2019/4/13.
12. Macro trend, https://www.macrotrends.net, last accessed 2019/4/13.
13. Butler Consultant, http://research.financial-projections.com/, last accessed 2019/4/13.

Lee, Minchul; Kulvatunyou, Boonserm.
"Design-for-Cost – An Approach for Distributed Manufacturing Cost Estimation."
Paper presented at APMS 2019 Conference, Advances in Production Management Systems, Austin, TX, United States. September 1, 2019 - September 5, 2019.

MODELING A SUPPLY CHAIN REFERENCE ONTOLOGY BASED ON A TOP-LEVEL ONTOLOGY

Farhad Ameri
Associate Professor
Engineering Informatics Lab
Texas State University
San Marcos, Texas, USA
ameri@txstate.edu

Boonserm Kulvatunyou
Research Scientist
Systems Integration Division
National Institute of Standards and Technologies (NIST)
Gaithersburg, Maryland, USA
serm@nist.gov

ABSTRACT

Several supply-chain ontologies have been introduced in the past decade with the promise of enabling supply chain interoperability. However, the existing supply-chain ontologies have several gaps with respect to completeness, logical consistency, domain accuracy, and the development approach. In this work, we propose a new, supply-chain, reference ontology that is 1) based on an existing top-level ontology and 2) developed using a collaborative, ontology-development, best practice. We chose this approach because empirical studies have shown the usefulness of adopting a top-level ontology both for improving the efficiency of the development process and enhancing the quality of the resulting ontology. The proposed proof-of-concept reference ontology is developed in the context of the Industrial Ontology Foundry (IOF). IOF is an international effort aimed at providing a coherent set of modular and publicly-available ontologies for the manufacturing sector. Although the proposed reference ontology is still at the draft stage, this paper shows that it has already benefited from the collaborative development process that involves inputs from the other working groups within IOF. Additionally, as a way to validate the proposed reference ontology, an application ontology related to a supplier discovery and evaluation use case is derived from the reference ontology and tested.

Keywords: **supply chain, reference ontology, manufacturing, collaborative ontology development, interoperability**

INTRODUCTION

Supply chains are increasingly more complex, digital, and dynamic. In this context, supply-chain integration is a necessary feature to enable enhanced coordination and communication among various supply chain participants such as vendors, service providers, and customers [1]. Many supply chain researchers and practitioners support the idea that supply chain efficiency can be improved with seamless flow of information [2]. One of the main enablers of such a seamless flow of information is interoperability. Interoperability is the ability of two or more systems to exchange information and interpret the exchanged information meaningfully and accurately in order to produce useful results via deference to a common information exchange reference model [3].

To date, supply chain interoperability is still a major, unsolved problem. The existing supply chain solutions have not been able to achieve full or agile information integration, because they do not interoperate [4]. Lack of interoperability can be attributed to differences in the underlying semantics and business rules implemented by different supply chain software systems.

Ontologies have been proposed as the solution to these differences. Simply put, an ontology, which is a controlled vocabulary represented by formal logic, provides a consensus-based set of terms for describing the types of entities in a given domain and the relations between them [5]. In the supply chain domain, the core entities include the organizations that form the supply chain, their internal functions, capabilities, and resources, the buying and selling processes, the materials and the information that flow throughout the supply chains, and the processes and services that govern the operation and coordination of the supply chain.

The first and most basic benefit of an ontology is that, like other kinds of standard data models such as entity-relationship model and XML Schema, it provides a common terminology that can be used for data annotation [6]. This common terminology enables both machines and humans to access, understand, search, and retrieve data more efficiently.

A secondary benefit of ontology stems from its logic-based nature. Unlike other kinds of data models, logically formulated ontologies allow human and machine agents to make *inferences* about operations such as data aggregation, comparison, querying, and quality assurance. In addition, when data models are annotated or tagged by ontological entities, they become more easily searchable, combinable, and analyzable using logical-reasoning implemented by compatible software tools.

These benefits are the main reasons that researchers have been proposing a growing number of supply chain ontologies [7-10]. Grubic and Fan [11] studied the existing supply-chain ontologies; they concluded that those ontologies have failed to

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

solve the current interoperability problems. The study identified several gaps in the existing supply chain ontology models [12]. Those gaps include weak methodological approaches, restricted and static views of supply chains, missing accounts of material traceability and service, and the dominance of taxonomies over formalized definitions.

A key conclusion from these studies is that "too much emphasis is placed on the organization and structure of human knowledge of supply chains rather than on understanding the ***reality*** of supply chains" (p.776). Ironically, the existence of these ontologies, which are based on varying and conflicting views of the supply-chain domain, has contributed to the interoperability problem rather than serving as a solution.

The objective of this current research is to investigate a method to develop a supply chain reference ontology based on a shared and domain independent, foundational ontology called Basic Formal Ontology (BFO). BFO's main difference from other top-level ontologies is its focus on reality (rather than an application or domain-specific view) and its past successes in using BFO in different domains. In the rest of the paper, we provide a background on BFO and on IOF where this research has been conducted. Then we outline the ontology development method. The draft supply chain reference ontology (SCRO) is presented, followed with a validation use case. Finally, the concluding remarks are provided.

INDUSTRIAL ONTOLOGIES FOUNDARY (IOF)

The IOF project is an international effort, with the participation of governments, industries, academia, and standard organizations. The vision of IOF is to make its ontologies publicly available and loyalty free in order to increase ontology adoptions in the manufacturing sector. The scope spans the entire domain of digital manufacturing in order to advance software and data interoperability.

The IOF results, once fully developed, will provide an open-source platform for developing, validating, aligning, sharing, and curating industrial ontologies. Rather than being an academic endeavor, IOF is committed to meet the needs of industrial stakeholders by providing reliable, turnkey solutions and by giving them best practices to integrate ontologies in their businesses. The technical goals of IOF include [12]:

- Create open, principles-based ontologies from which other domain-dependent or application-specific ontologies can be derived in a modular fashion.
- Ensure that IOF ontologies are non-proprietary and non-implementation-specific, so they can be reused in different industrial subdomains and standard bodies.
- Provide principles and best practices by which quality ontologies will support interoperability
- Institute a governance mechanism to maintain and promulgate the goals and principles.
- Provide an organizational framework and governance processes that ensure conformance to IOF principles and best practices.

Currently there are five active working groups (WGs) in IOF. Four of them addresses different subdomains of manufacturing including supply chain, production planning and scheduling, maintenance, and product-service systems. The last working group, namely the top-down WG, serves as the glue by providing a common ontology and ensuring the consistency across other working groups. The working groups receive support with respect to ontology development expertise from the members of the Technical Oversight Board (TOB) consisting of both ontology-inclined domain experts and ontologists [13]. Domain experts identify their interoperability requirements and ensure that the ontologists create definitions and axioms that meet those requirements.

Despite the broad scope of IOF, the ontologies it develops still must become a work item in existing standard development organizations (SDOs). One possible strategy is for IOF to develop detailed ontologies based on a few industry use cases. These detailed ontologies can then be modularized into mid-level reference ontologies and extended into domain-specific ontologies. IOF will maintain the mid-level reference ontologies; The SDOs will focus on the domain-specific ontologies There are two main concerns with this strategy: protecting all intellectual property rights and free access to the standards.

TOP-LEVEL ONTOLOGIES (TLO)

Ontologies can enable semantic interoperability when they are built according to a rigorous, multi-tiered, hierarchical architecture (Figure 1). Such an architecture has 1) a single, small, domain-neutral ontology at the top of this hierarchy and 2) a suite of lower-level ontologies – both domain-dependent and domain-specific ontologies. Top-level ontologies (a.k.a upper ontologies, or sometimes positioned in the opposite side as foundational ontologies) are highly abstract, domain-neutral because they establish a common framework for creating application ontologies [5].

An important function of an upper ontology is to support semantic interoperability by providing accurate and axiomatic definitions of the generic entities that can be further specialized by domain-specific ontologies. Some of the notable upper level ontologies include Basic Formal Ontology (BFO) [5], Domain Ontology for Linguistic and Cognitive Engineering (DOLCE) [14], PSL [15], and Suggested Upper Merged Ontology (SUMO) [16].

There are some differences in the granularity, structure, and philosophical underpinnings of upper-level ontologies. Nevertheless, several empirical studies have shown that using upper-level ontologies can improve both the quality and the efficiency of the ontology-development process [17]. In this research, BFO is investigated as the upper-level ontology. BFO has been used widely in the biological domain for integrating

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

disparate ontologies or developing interoperable ontologies for biological applications [19]. There are several reasons that make the investigation of using BFO as an upper ontology worthwhile for many domains including the supply chain domain. Firstly, BFO has a very large user base and it is widely used in a variety of ontologies including military and intelligence. Secondly, BFO is very small, with only 35 classes, and correspondingly easy to use and easy to learn. Additionally, BFO is very well-documented and there are multiple tutorials, guidelines, and web forums for using BFO in ontological projects.

As a domain-neutral upper-level ontology, BFO adopts a *realist approach* and represents different types of entities that exist in the world and relations between them. Realism-based ontologies are formalized descriptions that are based on scientific theories about the nature of entities in reality and the relationships between them. The notion of *ontological realism* amounts to the idea that an ontology should be analogous not to a data model, but rather to a reality model [18]. This maximizes the utility and stability of the ontologies that are based on BFO because a data model can be a specific view of the reality.

By choosing to investigate BFO first as the top-level ontology does not mean that it is necessarily the best top-level ontology for representing the domain of supply chain management. In fact, what we have observed so far in the IOF is that it is difficult, if not impossible, to single out one of the aforementioned top-level ontologies that can fully meet the requirements of the IOF working groups (much less the entire industrial subdomains) without workarounds or extraneous assumptions. For example, a question often arises to whether realism precludes the descriptions of non-existing or abstract entities such as a simulation model. While according to the authors of BFO, that is not the case; it is however a subject of validation within IOF. One of the objectives of IOF WGs is to experiment with multiple foundational ontologies and evaluate their strengths and weaknesses. We expect to conduct similar supply chain reference ontology modeling with other upper-level ontologies such as DOLCE in the future. At the time of writing this paper, IOF has not committed to adopting a single top-level ontology yet; and multiple scenarios, including allowing more than one top-level ontology, merging multiple top-level ontologies, or no top-level ontology, are currently being considered.

ONTOLOGY DEVELOPMENT METHODOLOGY

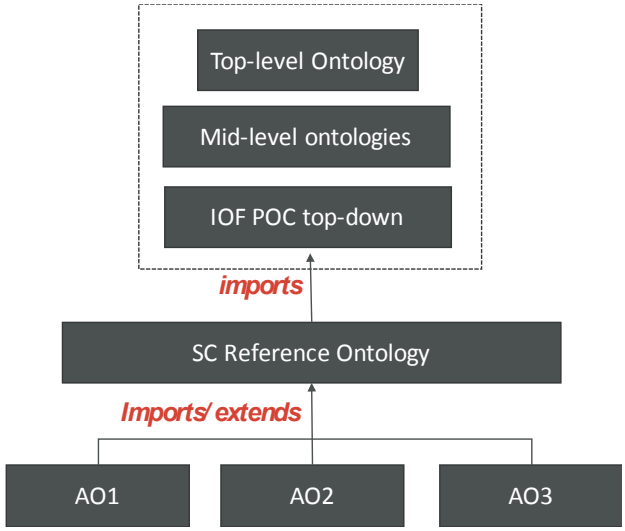
To develop the supply-chain reference ontology (SCRO), we imported BFO as the single, upper ontology. BFO splits all entities into two categories: *continuants* and *occurrents*. Continuants are the entities that continue to persist through time while maintaining their identity. Occurrents are the processes, events, or happenings in which continuant entities participate. Also imported are a few other domain-independent, mid-level ontologies. Examples of mid-level ontologies include time ontology, unit of measure ontology, and geospatial ontology.

The current version of SCRO uses Common Core Ontology (CCO), which bundles these multiple, mid-level ontologies. This approach conforms with the hub-and-spoke architecture recommended in the IOF's technical principle document [20]. SCRO also imports and extends the IOF proof-of-concept (POC), top-down ontology, which contains a small set of high-level terms such as *engineered system*, *product,* and *manufacturing resource*, that are common across multiple industrial manufacturing subdomains.

Different Application Ontologies (AO) in the domain of supply chain can import SCRO and further extend it to address application-specific requirements. The validation section in this paper describes how an application ontology related to the supplier discovery and evaluation is created based on the SCRO model. This tiered architecture is shown in Figure 1.

We adopted a bottom-up approach, driven by supply chain use cases, for developing SCRO. For this purpose, a template was designed for 1) describing the problem statement, 2) identifying the expected role of an ontology in the proposed use case, 3) listing the key notions (terms) related to the use case, and also 4) providing some Competency Questions (CQ) to be used later for validation purpose. Four use cases, related to different phases of supply chain, were proposed by the members to motivate the bottom-up development process. The proposed use cases were related to supplier discovery, supply-chain configuration, bidding automation, and supply-chain traceability. Table 1 shows the details of the supplier discovery use case.

The core terms related to different use cases were aggregated to create a draft list of terms composed of about 80 entries. The informal (natural language) definitions were created based on the procedure described in Figure 2. The informal definitions are human-readable definitions of the term and they are intended to be intelligible for subject-matter experts (SME).



Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

Figure 1. The position of SCRO in the stack of IOF ontologies

According to this process, the candidate definitions for each term are collected from internal sources (IOF members) and several external sources including Ontobee repository [21], ISO online browsing platform [22], relevant domain glossaries such as APICS [23]. If any of the collected definitions is deemed suitable to be used directly, it will be adopted as the informal (or Subject-Matter Expert- SME) definition as is. Otherwise, the candidate definitions will go through some linguistic and semantic pre-processing such as disambiguation, reconciling contradictions, removing unnecessary contextual contents, and removing redundancies to arrive at a more refined definition. Additional documentations were discussed in IOF including examples and corner cases.

Online sharable spreadsheets were the preliminary tool used for curating the list of terms and their definitions and also maintaining the history of their evolution. There are ongoing discussions to use better version-control systems, such as GitHub, and collaborative ontology development tools, such as Mobi [24]. Some examples of formal and informal definitions are provided in Table 2.

Table 1: Supplier Discovery use case

Problem Statement: *describe current state and future state*
Supplier discovery and search is often a manual, slow, and inefficient process. As the interaction between suppliers and customers becomes increasingly virtual and the lifespan of supply chains becomes shorter, more efficient and intelligent approaches to supplier search and evaluation are called for. One of the root causes of inefficiency in sourcing process is that manufacturing companies often publish and share their capabilities using informal and unstructured representation methods. Therefore, it is difficult to automate the sourcing or supply chain formation process.
How ontologies can help? (examples: search, data integration, decision support)
Primary utilities: • **Decision support/inference:** Ontologies can support human experts during sourcing process by providing inference-based answers to various queries about suppliers' capabilities.
Secondary utilities: • **Semantic Integration:** Ontologies can help with semantic integration of heterogeneous manufacturing capability models generated by dispersed actors. • **Automation:** Ontologies can enable machine agents to actively participate in supply chain formation process by proving machine-understandable content.
Competency Questions: (include at least 5 questions)
Which factories can machine complex geometries? What is the precision machining capability of this supply chain? What is the minimum wall thickness that can be machined in this factory? What is the largest diameter that can be turned in the factories owned by this company? What is the capability of this factory with respect to surface roughness? Does the capacity of this supply chain satisfy the demand?
Relevant terms:
• Supply chain • Supplier • Customer • Vendor • Manufacturing capability • Delivery lead time • Production capacity • Work order • Request for quote • Sourcing • Supply chain management • Inventory • Resource utilization • Supply • Demand • Seasonal demand • Factory • Machine Cell • Capacity • Transportation mode

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

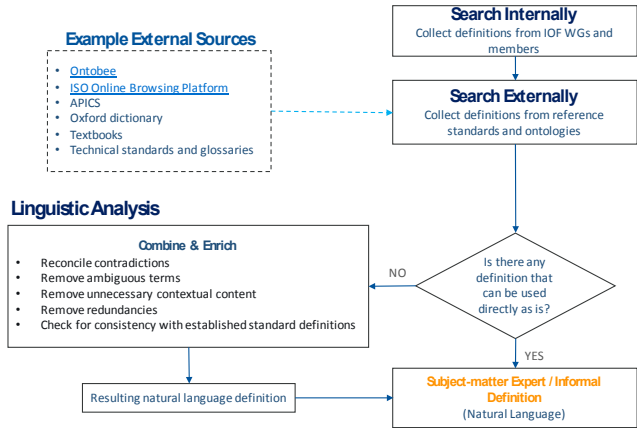


Figure 2. The overall process of creating subject-matter expert definition

Once a consensus is reached on the informal definition of a term, the ontological analysis begins by arranging the terms into a hierarchy based on their BFO or IOF classes. Then formal definitions of the terms are created. Formal definitions should use the vocabulary of the relevant upper-tier ontologies and follow the Aristotelian definition strategy, whereby each definition is of the form 'A =def. a B which Cs'. B is the parent term/class of A in the ontology, and C is the *differentia* which marks out those Bs which are As. The last step in ontological analysis is adding formal axioms (such as equivalence and sub-class axioms) to each term to create the formal expression of the term in OWL language. This process results in creation of ontological classes or universals.

SCRO Requirements:

The specific requirements for SCRO are listed below. The terms "SHOULD" and "SHALL" are to be interpreted as described in RFC2119:

- SCRO SHOULD be small and modular
- All classes in SCRO SHALL be subclasses of a class in the mid-level or top-level ontology.
- SCRO SHALL reuse the existing relationships available in mid-level and top-level ontologies to the extent possible.
- SCRO SHOULD provide the necessary generic classes that can be used for representation of different supply chain processes, roles, functions, capabilities, material entities, and information entities.
- SCRO SHOULD provide the necessary building blocks for developing supply chain application ontologies with strategic, tactical, and operation focus.
- SCRO SHOULD be sufficiently axiomatized to enable interoperability and application integration.

Table 2: Examples of formal and informal definitions of SCRO terms (CCO stands for Common Core Ontologies. The prefix indicates the source ontology for an imported term)

Term	Formal Definition	Informal Definition
Supply Chain	A CCO:GroupOfOrganizations involved in trading IOF:Products and IOF:Services and other business relationships with one another.	supply chain is a set of companies and other organizations involved in trading and other business relationships with one another.
Supplier	An IOF:Organization or IOF:Person with a IOF:SupplierRole	An organization or person who sells products or services.
Supplier Role	An IOF:SupplierRole is a BFO:Role inhering in anCCO:Agent that, if realized, is realized in some act of selling.	[no SME definition necessary since it is a construction entity and not a user-facing entity]

SUPPLY CHAIN REFERENCE ONTOLOGY

The main elements of a supply chain are the organizations that comprise the supply chain. Other elements include 1) the materials and information that flow throughout the supply chain and 2) the processes in which those material and information entities participate. This section describes how the aforementioned entities are further formalized ontologically in SCRO through definitions and axioms. These axioms are in draft statuses. First-Order Logic (FOL) notation is used for axioms shown in this paper. In FOL notation, → denotes a sub-class axiom and ≡ denotes an equivalence class axiom.

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

Supply Chain Material Entities

In BFO, one of the most relevant subclasses of continuants is the *material entity* class. It is a continuant that includes some portions of matter as part. Machines, physical products, raw materials, people, and organizations are examples of material entities in the domain of supply chain. Selected axioms related to organization, person, agent, supplier, customer, and manufacturing resource are illustrated in this section. Some of the presented axioms may appear to be too weak or too strong in an absolute sense. We expect that as SCRO is validated by more application ontologies, the axioms will be adjusted to meet the formalization requirement of different use cases.

organization is a subclass of bfo:object aggregate.

organization(x) → object-aggregate(x).

person subclass of bfo:object.

person(x) → object(x).

Every **agent** at time t is a **person** or **organization**.

instance-of(x,agent,t) ≡ (instance-of(x,person,t) or instance-of(x,organization,t)).

supplier and **customer** are subclasses of agent

supplier(x) → agent(x).

customer(x) → agent(x).

The axioms related to *role classes* usually contain the notion of time (t) because entities can bear different roles in different time intervals.

A **supplier** is an agent who bears a supplier role.

Instance-of-supplier(x, supplier,t) ≡ ∃y(supplier-role(y) & has-role(x,y,t)).

A **manufacturing supply chain** is a group of suppliers connected by supply chain links to manufacture a certain product.

A **supply chain link** (a sub-property of *participates-in*) is a relationship between two suppliers (s1 and s2) if s participates in the process of supplying material and information to s2.
The axiom that can be used for querying if a supplier (s1) is a member of a supply chain (sc1) is as follows:

supplier (s_1) & supply chain (sc_1) & (∃ s_2(supplier(s_2) & has-supply-chain-link (s_1, s_2) & member-of (s_2, sc_1)) → member-of (s_1, sc_1)

manufacturing resource is a bfo:continuant that bears a manufacturing resource role.

instance-of (x,manufacturing resource,t) ≡ ∃y(manufacturing resource role(y) & has-role(x,y,t)).

A **piece of manufacturing equipment** is a piece of equipment that bears a function and any process which realizes that function is a manufacturing process.

piece of manufacturing equipment(x) ≡ piece of equipment(x) & ∃f(has-function(x,f) & ∀p(process(p) & realizes(p,f))→ manufacturing process(p)).

Supply Chain Roles

The *role* class in BFO provides a versatile template for creating different *defined classes* when an entity is in some special natural, social, or institutional set of circumstances. Role is a *realizable entity* in BFO, meaning that it is only manifested or realized in certain conditions and certain times. Examples of supply chain roles include 1) transportation role that is the role of an organization to serve as a provider of transportation service in a supply chain or 2) product role that is the role of an artifact to be sold or exchanged in a supply chain. Role may also be viewed as a work around to maintain a single inheritance hierarchy. In other words, an entity is born or created or designed to be what it is (the class in which is asserted to a member) but can engage in many roles. The **supply chain member role** is used as the top class for different types of supply chain roles such as material provider role or test service provider role. Figure 3 shows some of the sub-types of supply-chain-member role. The formal definitions of important supply chain roles, together with their axioms, are provided below.

supplier role subclass of role.

supplier-role(x) → role(x).

A **supplier role** is a role inhering in an agent that, if realized, is realized in some act of selling.

supplier-role(x) ≡ ∃y(agent(y) & has-role(y,x) & ∃p(process(p) & realizes(y,p) → act-of-selling(p))).

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

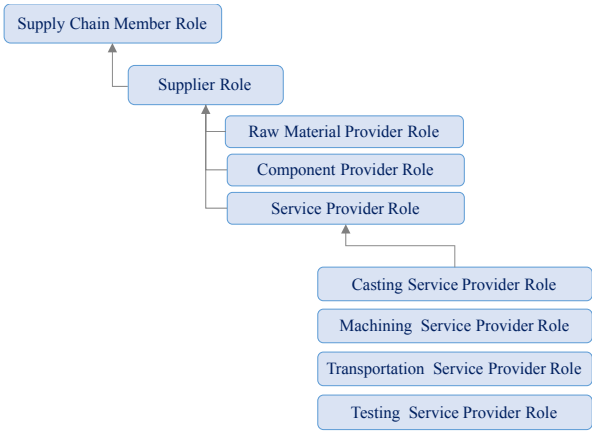


Figure 3. Different sub-classes of supply chain member role class

Note that the supplier role is only related to the *act of selling* (a product or a service). There is no mention of the particular product that will be produced as a result of the participation of the supplier. There are three different sub-types of supplier role in SCRO, namely, raw material provider role, component provider role, and service provider role.

A **service provider role** is a role inhering in an Agent that is realized in some act of service provision. A **machining service provider role** is a role inhering in an agent that, if realized, is realized in some act providing a machining service.

Machining-service-provider- role(x) ≡ ∃y(agent(y) & has-role(y,x) & ∀p(process(p) & realizes(p,x) → act-of-providing-machining-service(p)).

Accordingly, a **machining supplier** is a supplier that bears a machining service provider role.

machining supplier(x) ≡ supplier (x) &(∃y(machining service provider role(y) & has-role(x,y)).

The **customer role** is the role inhering in an agent in a basic economic transaction from the point when a purchasing act has been initiated through to completion.

instance-of(x,customer-role,t) ≡ ∃y,z(agent(y,t) & has-role(y,x,t) & ∃w(instance-of(w,act-of-purchasing,t) & agent-in(y,w,t))).

Figure 4 shows the function, capability, and the role related to a *Machining Supplier*. In this figure, the green boxes are BFO classes, while the blue boxes denote SCRO classes. The white boxes represent the classes imported from mid-level ontologies such as IOF top-down ontology or CCO. Figure 5 shows the class structure of some of the SCRO continuants.

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

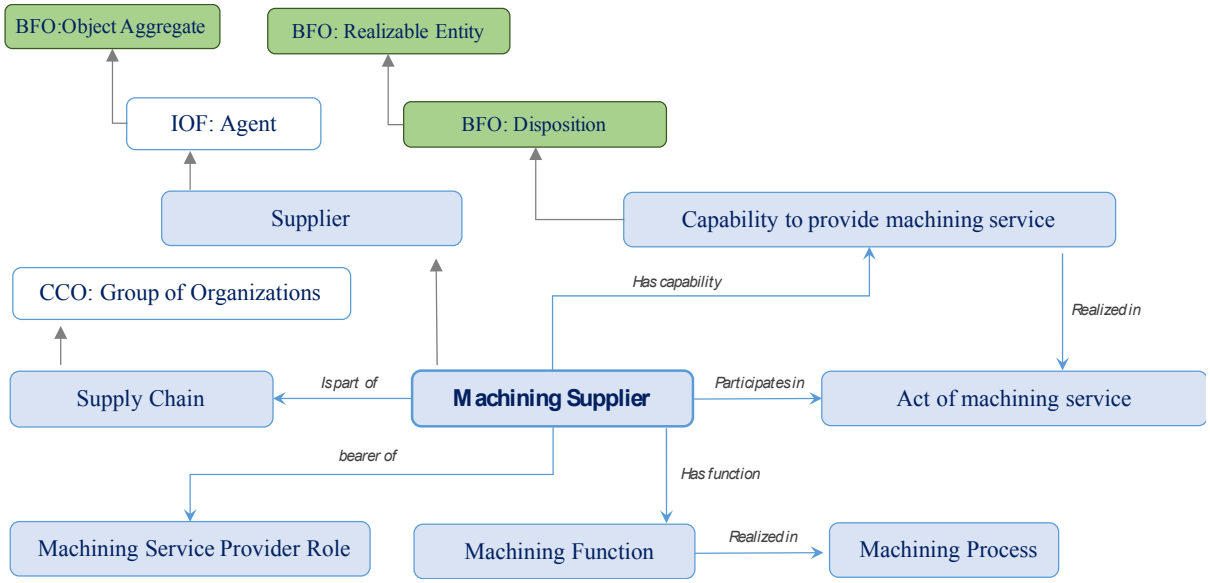


Figure 4. The class diagram for the Machining Supplier class

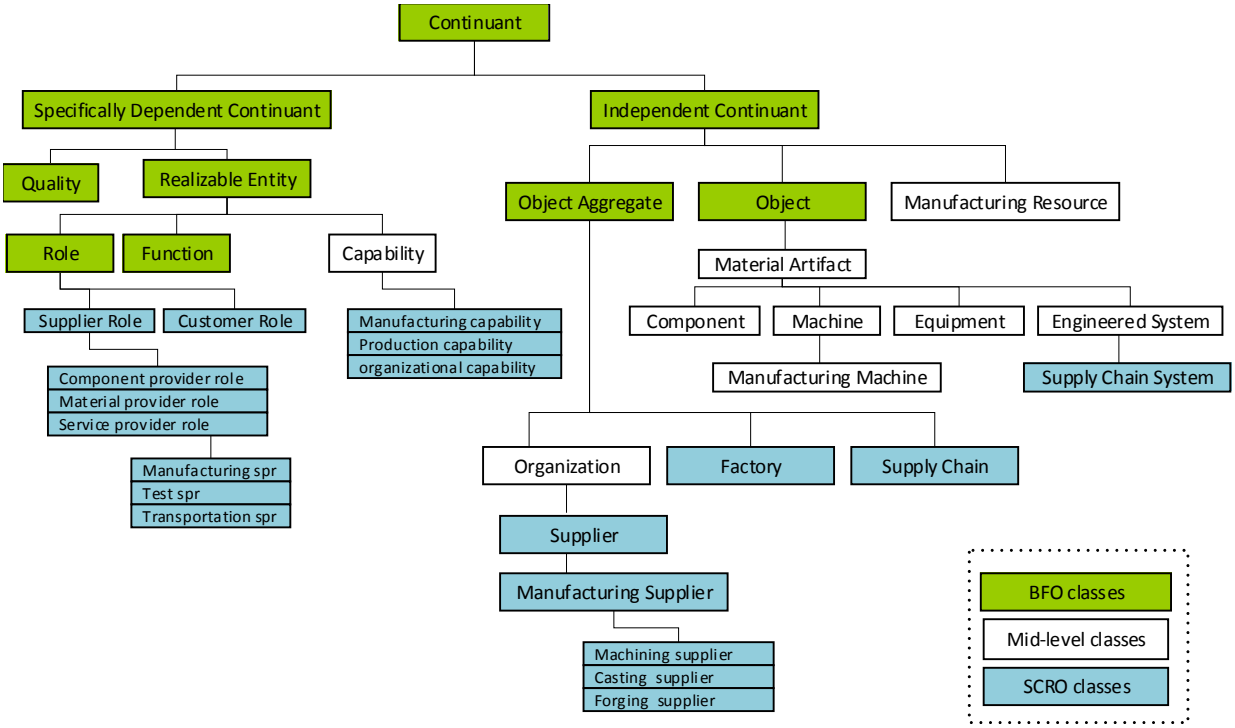


Figure 5. The continuant side of SCRO (partial view)

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

Network Representation of Supply Chains

Supply chains can be perceived as a network in which the nodes represent the partnering organizations and the edges denote the flow of material and information between the organizations (Figure 6).

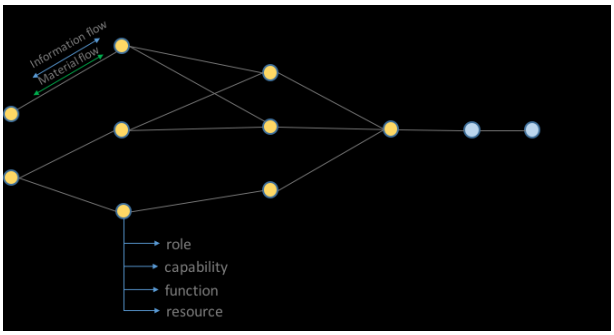


Figure 6. Network representation of supply chain

The organizations within a supply chain have different roles, functions, capabilities, and resources. SCRO, as a reference ontology, provides the necessary building blocks for modeling supply chains as a network. That is, the supply chain link, that was defined in the previous section, represents an edge in the network. A particular supply chain then can be defined as a group of organizations that participate in production of a specific product. The material entities, such as different types of raw materials, products, and semi-finished assemblies, that flow through the network have their own ontological representation.

As shown in

Figure 7, the flow of material between two suppliers is modeled as a participation of a material entity in an *Act of Shipment*. Act of Shipment (an Occurrent) is a sub-class of Act of Location Change in Common Core Ontology (CCO). An Act of Location Change is defined as *An Act of Motion*, in which the location of an Object is changed by some Agents. Supplier 1 (S1) participates in the act of shipment as the sender of the material entities and Supplier 2 (S2) participates in this act as the receiver of those material entities sent by S1.

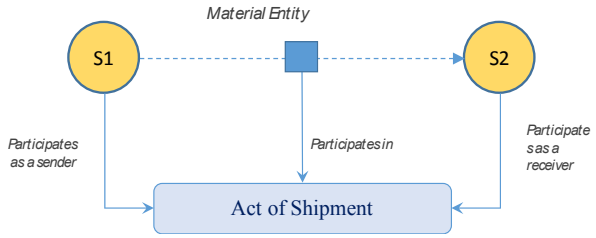


Figure 7. Flow of materials between two nodes of the supply chain: supplier 1 (S1) and supplier 2 (S2).

Using a similar approach, information communication can be modeled as an *Act of Communication* in with the participation of a sender (S1) and a recipient (S2) participate. The Information Entity itself is one the participants in the act of communication.

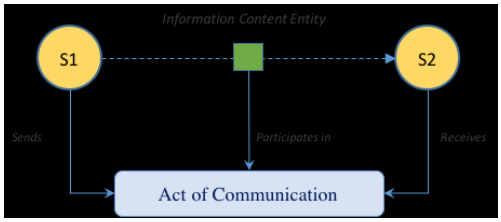


Figure 8. Flow of information entities between two nodes of the supply chain: supplier 1 (S1) and supplier 2 (S2).

The definitions for *sends* (information) and *receives* (information) properties are adopted from Common Core Ontology:

- **sends** (*inverse of* **has sender**) is a relationship between an Agent *a1* and an Act Of Communication *c1* and such that *a1 sends c1* if and only if *a1* is the initiator and encoder of the Information Content Entity participating in *c1*.
- **Receives (***inverse of* **has recipient)** is a relationship between an Agent *a1* and an Act Of Communication *c1* such that *a1 receives c1* if and only if *a1* is the **recipient and decoder** of the Information Content Entity participating in c1.

When a network model of the supply chain is available, interesting information can be derived through reasoning and querying the network in order to answer important business questions (competency questions) such as:

- What are the outputs of this supplier in this supply network?
- What types of information does this supplier receive?

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

- What is the most connected supplier in this supply network?
- What are the first-tier/second-tier suppliers in this network?
- Which supplier provides raw materials in this network?
- What is the path followed by a certain component in this network?

SUPPLIER DISCOVERY USE CASE

To validate the proposed reference ontology, we focused on a supplier sourcing use case. Various software agents collaborate and interact with the end goal of forming a supply chain for a given production order. Consequently, interoperability is an essential feature of this use case since the agents operate independently as autonomous entities. For the purpose of this use case, SCRO was extended to create an *application ontology* tailored based on the needs of an agent-based sourcing scenario. The participating machine agents in this use case provide different types of web services as described below:

- **Capability Advertisement Service (CAS):** Each manufacturing company is represented by a CAS agent which responds to queries regarding the manufacturing capabilities of the company. The capabilities are modeled using the extensions of the "capability" module of SCRO. In this use case, it is assumed that the capability data is coded as ontology instances. Another plausible scenario is to keep the capability data in data structures that are tagged using ontological entities.
- **Supplier Evaluation Service (SES):** The Supplier Evaluation Service evaluates suppliers with respect to their abilities in fulfilling the requirements of specific production orders. SES receive requests from the Supply Chain Configuration agents. Different SES agents may use different methods and algorithms for supplier evaluation.
- **Supply Chain Configuration Service (SCCS):** SCCS agents are in charge of building a supply chain that can complete a production order and manufacture the requested components or assemblies in the requested volumes and delivery times per specifications. They interact and communicate with SES agents to identify qualified suppliers that can participate in the desired supply chain.
- **Supply Chain Evaluation Service (SCES):** SCES agent evaluates the performance of a supply chain using key performance measures such as reliability, responsiveness, and cost.

In this use case, an ontology can serve two purposes: 1) to enable interoperability among heterogenous agents and 2) to enable capability analysis and inference based on the explicitly stated capabilities. Figure 8 shows how the subclasses of the *capability* class are extended for this use case. SCRO contains three types of capability classes, namely, manufacturing capability, production capability, and organizational capability. The manufacturing capability class was extended to represent the classes needed for asserting the capabilities related to attributes of manufactured artifacts. For example, *part material capability* is related to the capabilities of a certain organization to process different types of materials. Similarly, *part complexity capability* is the class related to range of geometric complexities that can be accommodated by a manufacturer.

Capability Representation

The CAS agent can represent and advertise the capabilities of a manufacturing company both directly and indirectly. Direct capability representation entails providing values for different measures of manufacturing capability such as complexity, material, and precision capability measures. Figure 10 shows the template used for expressing the value of surface roughness capability. According to this template, *Part Surface Roughness Capability* is a capability that is measured as a *length measurement datum* which is a type of *Measurement Datum*, a class imported from the Information Artifact Ontology (IAO).

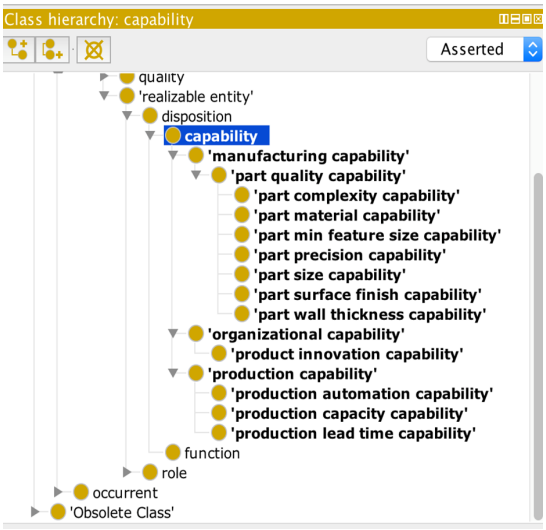


Figure 9: The extended capability class

Indirect representation of capabilities involves two steps. First, we describe the manufacturing resources owned by the company. Second, we allow the supplier-evaluation agents to interpret the capabilities based on the available resources. The capability inference methods described in the next section are based on indirect capability descriptions. In doing so, we assume that the company instantiates the *factory* class, which is considered to be an *object aggregate* in SCRO. The manufacturing resources are linked to the factory class using *has part* relationship.

In many occasions, capability-related queries can be formulated as SPARQL queries. For example, the query shown

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

in Figure 11 returns all factories that can provide vertical-milling services for parts that are longer than 10 inches in length. However, more sophisticated reasoning might be needed when, for example, the supplier-evaluation agent is interested in learning about different ranges of part complexities that can be supported by a given supplier. In this example query, MSD [26] is imported as an external ontology to enhance the expressiveness of the application ontology which is developed through extending SCRO.

Capability Inference

One of the functions of the Supplier Evaluation Agent is capability inference. Using the extended capability module of SCOR, one can create a formal representation of a factory. From the capabilities explicitly represented in the factory model, new capabilities can be inferred automatically using the ontological reasoning services. Four categories of capability are discussed in this section: 1) surface roughness capability, 2) process capability, 3) material capability, and 4) production capability. It should be noted that such inferences at best provide an approximation of the latent capabilities of a supplier.

Figure 10: Surface roughness capability measurement template

Active Ontology × | Entities × | Individuals by class × | DL Query × | SPARQL Query ×

SPARQL query:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX msdl: <http://infoneer.txstate.edu/ontology/>

SELECT ?factory_instance
WHERE
{
   ?factory_instance a msdl:factory .
   ?vertical_milling_capability_instance a msdl:"vertical milling capability" .
   ?factory_instance msdl:MSDL_0000768 ?vertical_milling_capability_instance .
   ?length_measurement_datum_instance a obo:IAO_0000408 .
   ?length_measurement_datum_instance obo:IAO_0000004 ?length_value .
   ?part_length_capability_instance a msdl:MSDL_0000674 . #
   ?part_length_capability_instance msdl:MSDL_0000924 ?length_measurement_datum_instance .
   ?production_machine_instance a/rdfs:subClassOf* msdl:MSDL_0000033 . # "
   ?production_machine_instance msdl:MSDL_0000936 ?part_length_capability_instance .
   ?factory_instance msdl:MSDL_0000129 ?production_machine_instance .
    FILTER(?length_value > 10)  }
```

Figure 11: A SPARQL query returning all factories with vertical milling capabilities for parts longer than 10 inches

Surface finish capability:

A machine tool can create certain qualities such as tolerance, surface roughness, or minimum feature size on a part. The range of these qualities define the capability of the machine tool. The collective capability of the factory is calculated by aggregating the capabilities of individual machines in the factory. Figure 12 shows the procedure for calculating the surface-finish capability that can be provided by a factory operated by a given company. According to this procedure, each machine's surface-finish. capability value, which is already stored as instance information, is retrieved first. If the retrieved value is null, then the immediate superclass of the machine tool is queried instead and the surface -finish capability value is retrieved. The reason behind this approach is that the parent type of machine class can provide a reasonable approximation of the capabilities of the children types of machines. If none of the higher-level individuals can provide a value for surface- finish capability, then a similar, generic machine from the same machine vendor is used to provide some approximation about the capability of the machine. The generic machine from a given vendor is the average machine with respect to capabilities based on the vendor's product portfolio.

FOR i=1 to num of machines in the factory f [instance of SCRO: factory]
m_i [instance of SCRO: machine tool]
Retrieve **part surface finish capability** value m_i_ sf_{cap}
IF m_i- sf_{cap} = Null
    Find an instance of the superclass of m_i => sm_i
Retrieve **part surface finish capability** value sm_i_ sf_{cap}
    IF sm_i _ sf_{cap} = null THEN
    Find the instance of a **Generic Machine** from the same vendor => gm_i
        Retrieve **part surface finish capability** value gm_i _ sf_{cap}
            IF gm_i- sf_{cap} = Null THEN Let m_i- sf_{cap} = null
                ELSE Let m_i- sf_{cap} = gm_i- sf_{cap} AND go to the next machine
                End IF
        ELSE Let m_i- sf_{cap} = sm_i _ sf_{cap} AND go to the next machine
        End IF
End IF
Factory - sf_{cap} =Min (m_i- sf_{cap}) for all values of i

Figure 12: The procedure of calculating the surface finish capability of a factory

This procedure is based on the simplifying assumption that surface-finish capability is a standalone capability. However, more realistically, surface-finish capability is related to other types of capability such as surface-area capability or material capability.

Material Capability:

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

Material capability is also inferred based on the *is-a* relationship between different instances of materials. The factory model contains a list of materials that can be processed by the factory. The ontological reasoner can identify all super-classes of the explicitly stated material types. The instances of the identified upper- level classes are then added to list of materials that can be processed at the factory as inferred materials.

The logic behind this approach is that if a particular vertical mill, for example, can machine a special grade of aluminum then it can also machine more generic grades of aluminum as well. Material capability, in most real-life scenarios, is evaluated in relation with other capabilities. For example, the grade of material may impact the achievable tolerances and surface finishes on a given machine tool. These dependencies between capabilities can be encoded in the ontology through defining semantic rule which is outside the scope of this paper.

Process Capability:

A manufacturing process is the realization of an explicit manufacturing function intended for a piece of manufacturing equipment. When instances of manufacturing equipment are added to the factory, the manufacturing functions associated with the equipment are added to the list of available functions in the factory. The functions added directly through the equipment are considered to be explicit functions. The ontological reasoner identifies all sub-classes of the explicit functions as the inferred functions. For example, if a machine in a factory has a 'turning' function, then all instances of sub-classes of turning (including boring, facing, grooving, threading) are added to the list of 'inferred' processes (functions) for that factory.

Production Capability:

Production capability of a manufacturing facility is related to factors such as the number of equipment and the variety of the products that can be produced. Simplistically, the capacity of the factory directly depends on the bottleneck resource in the factory. There are some other indirect factors, such as the availability of the preventive maintenance system that can alter the capacity.

The capacity-capability class can be measured as an ordinal measurement datum with low, medium, high values. The variety capability, also measured as an ordinal measurement datum, depends on the available types and variety of manufacturing processes. Availability of more manufacturing functions can imply a higher variety of manufacturable parts. Therefore, the reasoner should consider both the explicit and the inferred processes when calculating the variety capability of a factory.

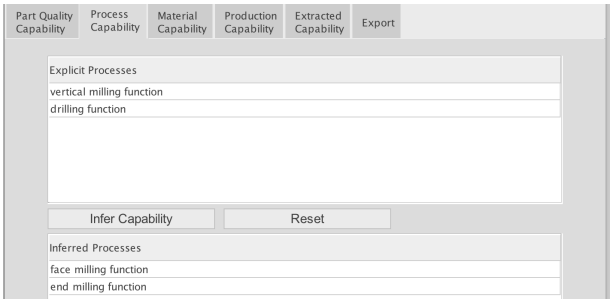


Figure 13: Proof-of-concept tool for capability inference

Based on the described procedures, a proof-of-concept tool was developed for 1) analyzing different capabilities of manufacturing suppliers and 2) inferring new capabilities based on the explicit capabilities. Figure 13 shows one of the user interfaces of the developed tool related to process capability analysis.

CLOSING REMARKS

In this paper, we presented the initial version of a reference ontology (SCRO) that can be used for creating more specialized application ontologies for the supply-chain domain. SCRO was developed based on the methodologies and specifications recommended by IOF. BFO was used as the top-level ontology. One of the objectives of the current experiment was to evaluate BFO as an upper ontology for the supply-chain domain. Different criteria such as ontological completeness, logical consistency, and accuracy in capturing the domain were used to perform that evaluation.

The results of the evaluation indicated that the breadth of coverage of BFO is adequate to represent commonly used entities in the supply chain domain. Also, the underlying axioms of the top-level ontology provided the needed logical consistency for the investigated use cases. However, further testing is needed to evaluate the adequacy of BFO for more complicated use cases, in which interoperability and applications integration are the main concerns.

One of the issues related to BFO is its inaptitude to represent the abstract notions that is rooted in its realist approach. Examples of abstract notions in supply-chain domain include the various supply chain-models that do not exist but need to be represented; e.g., for simulation and analysis. Workarounds are often needed to address abstract notions in BFO. In the future, addtional application ontologies will be created based on SCRO to evaluate its adequacy for different use cases that require interoperability and inference services.

DISCLAIMER AND ACKNOWLEDGEMENT

Certain commercial systems and applications identified in this paper are not intended to imply recommendation or

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

endorsement by the National Institute of Standards and Technologies, nor is it intended to imply that they are necessarily the best available for the purpose. We acknowledge the input we received from the IOF community.

REFERENCES

[1] Kim, H. M., and Laskowski, M., 2018, "Toward an ontology-driven blockchain design for supply-chain provenance," Intelligent Systems in Accounting Finance & Management, 25(1), pp. 18-27.

[2] Tian, K. S. a. Y., "Supply Chains Integration: Architecture and Enabling Technologies," Journal of Computer Information Systems(- 3), pp. - 67.

[3] Chapurlat, V., and Daclin, N., 2012, "System interoperability: definition and proposition of interface model in MBSE Context," IFAC Proceedings Volumes, 45(6), pp. 1523-1528.

[4] Rayyaan, R., Wang, Y., and Kennon, R., 2014, "Ontology-based interoperability solutions for textile supply chain," Advances in Manufacturing, 2(2), pp. 97-105.

[5] Arp, R., Smith, B., and Spear, A. D., 2015, Building Ontologies with Basic Formal Ontology, The MIT Press.

[6] Guarino, N., Oberle, D., and Staab, S., 2009, "What Is an Ontology?," pp. 1-17.

[7] Fox, M. S., Barbuceanu, M., and Gruninger, M., 1996, "An organisation ontology for enterprise modeling: Preliminary concepts for linking structure and behaviour," Computers in Industry, 29(1), pp. 123-134.

[8] Soares, A. L., Azevedo, A. L., and de Sousa, J. P., 2000, "Distributed planning and control systems for the virtual enterprise: Organizational requirements and development life-cycle," Journal of Intelligent Manufacturing, 11(3), pp. 253-270.

[9] Scheuermann, A., and Leukel, J., 2014, "Supply chain management ontology from an ontology engineering perspective," Computers in Industry, 65(6), pp. 913-923.

[10] Geerts, G. L., and O'Leary, D. E., 2014, "A supply chain of things: The EAGLET ontology for highly visible supply chains," Decision Support Systems, 63, pp. 3-22.

[11] Grubic, T., and Fan, I.-S., 2010, "- Supply chain ontology: Review, analysis and synthesis," Computers in Industry, 61(-8), pp. 776-786.

[12] 2019, "IOF Charter," https://www.industrialontologies.org/iof-charter/.

[13] Kulvatunyou, B., Wallace, E., Kiritsis, D., Smith, B., and Will, C., 2018, "The Industrial Ontologies Foundry Proof-of-Concept Project," IFIP WG 5.7 International Conference, APMS 2018, Springer, Seoul, South Korea.

[14] Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, R., Schneider, L., and Partner Istc-cnr, L., 2002, WonderWeb Deliverable D17. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology.

[15] Gruninger, M., and Menzel, C., 2003, "The Process Specification Language (PSL) Theory and Applications," AI Magazine, 3(24).

[16] Niles, I., and Pease, A., 2001, "Towards a standard upper ontology," Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001, ACM, Ogunquit, Maine, USA, pp. 2-9.

[17] Keet, C. M., "The Use of Foundational Ontologies in Ontology Development: An Empirical Assessment," Proc. The Semantic Web: Research and Applications, G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and J. Pan, eds., Springer Berlin Heidelberg, pp. 321-335.

[18] Smith, B., and Ceusters, W., 2010, "Ontological realism: A methodology for coordinated evolution of scientific ontologies," Applied Ontology, 5(3-4), pp. 139-188.

[19] Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V., 2015, "The role of ontologies in biological and biomedical research: a functional perspective," Briefings in Bioinformatics, 16(6), pp. 1069-1080.

[20] 2019, "IOF Technical Principles Document," https://www.industrialontologies.org/iof-technical-principles-document/.

[21] Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J., Mungall, C., Courtot, M., Ruttenberg, A., and He, Y., 2017, "Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration," Nucleic acids research, 45(D1), pp. D347-D352.

[22] 2019, "ISO Online Browsing Portal (OBP)," https://www.iso.org/obp/ui/.

[23] Jr., J. H. B., 2013, "APICS Dictionary," APICS, Chicago, IL.

[24] 2019, "Mobi," https://mobi.inovexcorp.com.

[25] Smith, B., and Ceusters, W., 2015, Aboutness: Towards Foundations for the Information Artifact Ontology.

[26] Ameri, F., and Dutta, D., "An upper ontology for manufacturing service description," Proc. 2006 ASME International Design Engineering Technical Conferences and Computers and Information In Engineering Conference, DETC2006, September 10, 2006 - September 13, 2006, American Society of Mechanical Engineers.

Kulvatunyou, Boonserm; Ameri, Farhad.
"Modeling a Supply Chain Reference Ontology Based on a Top-Level Ontology."
Paper presented at ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019, Anaheim, CA, United States. August 18, 2019 - August 21, 2019.

Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data

Madhusudanan Navinchandran[1], Michael E. Sharp[2], Michael P. Brundage[3], and Thurston B. Sexton[4]

[1,2,3,4] *Systems Integration Division, Engineering Laboratory,*
National Institute of Standards and Technology
Gaithersburg, MD 20899
madhusudanan@nist.gov
michael.sharp@nist.gov
michael.brundage@nist.gov
thurston.sexton@nist.gov

ABSTRACT

Maintenance Work Orders (MWOs) are a useful way of recording semi-structured information regarding maintenance activities in a factory or other industrial setting. Analysis of these MWOs could provide valuable insights regarding the many facets of reliability, maintenance, and planning. Information such as which maintenance activities consume the most work hours, identification of problem machines, and spare parts needs can all be inferred to some degree from well-documented MWOs. However, before one can derive insights, it is first necessary to transform the data in the MWOs (generally some form of natural language) into something more suitable for computer analysis. The National Institute of Standards and Technology (NIST) previously developed a computer aided tagging system that allows for the quick identification of key concepts within the natural language of the MWOs, and a protocol for categorizing these concepts as *problems*, *solutions*, or *items*. Using this annotation method, this paper investigates machine learning methods to gain insights about work hours needed for various maintenance activities. Through these techniques, it is possible to explain the factors captured in the MWOs that have the strongest relationship with the duration of maintenance actions. The workflow of this research is to first build strong data-driven models to classify the duration of any maintenance activity based on the language and concepts gathered from the associated MWO. Sensitivity analysis of the inputs to these classifiers can then be used to determine relationships and factors influencing maintenance activities. This paper investigates two machine learning models - a neural network classifier and a decision tree classifier. Input features for the classifier were the annotated concept tags for solutions, problems and items derived from MWOs of an actual manufacturer. This process for gaining insights can be generalized to various applications in the maintenance and Prognostics and Health Management (PHM) communities.

1. INTRODUCTION AND BACKGROUND

Optimizing maintenance activities is important throughout nearly every industrial setting, such as, manufacturing, chemical plants, process engineering (e.g., nuclear plants), and operations of field equipment (e.g., construction and mining equipment). Large scale implementations of maintenance have evolved from primarily reactive maintenance, to regular preventive maintenance scheduling (Sherwin, 2000), then on to more condition based and predictive maintenance practices (Camci, 2015; Helu & Weiss, 2016). Although the dependence and role of the human practitioners has evolved as well, much of the importance of workers has remained the same. The maintenance technician remains an important sensing tool and decision maker in observing symptoms, diagnosing issues, as well as prescribing and enacting maintenance activities. Such specialized, experience-based knowledge is often termed as *tribal* knowledge (Allen, 2013). Some of this knowledge is captured as the natural language put into Maintenance Work Orders (MWOs). Typically, information in an MWO is manually populated by an observing technician when a problem is faced at an asset, conveying details about how the issue was diagnosed and how it was resolved at each stage of the maintenance process. Such MWOs represent a wealth of useful knowledge about the sequence, description, causality, and timing of events with respect to the problem and its resolution. However, given the heavy involvement of manual free-form natural language, this data is often very inconsistent and inadequate for direct computer based analysis. The nature of natural language is such that variations will be

Madhusudanan Navinchandran, Fnu; Sharp, Michael; Brundage, Michael; Sexton, Thurston.
"Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

present that require some form of cleaning, translation, and consolidation in preparation for computer based analytics. This paper utilizes a previously developed method to clean and prepare textual data from MWOs (Sexton, Brundage, Hoffman, & Morris, 2017), and then proceeds with its analysis.

By providing methods to analyze this data, knowledge in MWOs can be captured and used to address potential trouble spots (e.g., most problematic machines) and prioritize actions, such as scheduling of operations including maintenance. Particularly in context of this work, models of maintenance action duration derived from MWOs can assist in maintenance action scheduling and determination of anomalous instances that may highlight additional underlying problems.

This paper relates concepts and ideas discovered in natural language field entries of MWOs to the expected duration of the action set associated with that work order. Often the language used to describe an action can give some indication of the relative severity or commitment of resources needed to complete that action. For example, *replacing* a part may tend to consume more time than a simple *adjustment* of a corresponding part. However, following the nature of natural language, there are no absolutes in regard to single words or qualifying concepts. Instead, concepts and ideas are treated as 'soft' influencing factors that may result in different durations given context. Consider that *replacing* lubricant oil consumes less time than *replacing* a motor - hence each solution might have a unique distribution of associated durations. To model the relations between MWO terms and maintenance time, machine learning models are studied and suitably adapted in this paper.

The objectives for this study are to,

- categorize various maintenance actions into meaningful groups based on activity type and duration
- identify the most influential features of maintenance, such as important *problems*, *solutions* and physical *items*
- identify outlier MWO instances to trigger deeper root cause analysis investigations.

The rest of the paper is structured as follows. The state of the art for natural language analysis in manufacturing and current time metrics in maintenance is covered in Section 2. Research challenges and overall methodology for carrying out the study is described in Section 3. A use case showing the implementation of the methodology on a manufacturing MWO dataset is demonstrated in Section 4. Results and issues from the analysis, and scope for extending this research are discussed in depth in Section 5, and conclusions are presented in Section 6.

2. CURRENT STATE OF THE ART

The relationship between maintenance and time related data is often discussed in the domains of reliability and performance monitoring. For example, literature discusses modeling for maintenance intervals with respect to time, by using distributions such as Gaussian, Weibull or Gamma (Locks, 1973). Additionally, there are time-related metrics that focus more directly on the equipment quality and performance, such as the Mean Time Between Failures (MTBF) (Gulati & Smith, 2009). Some research even investigates comparing performance of time-based (calendar-based maintenance) versus condition-based maintenance techniques using only failure time data sets (Ahmad & Kamaruddin, 2012). However, rarely in literature is the time taken (duration) for specific maintenance actions investigated as it relates to those activities, and in particular utilizing information gained from MWOs.

Mukherjee and Chakraborty (2007) discuss work towards understanding the text of maintenance logs to gain insights, such as diagnostic fault trees, from these logs. As mentioned in Section 1, the process of extracting this type of information from MWOs data is not straightforward. The difficulty in processing the language is often due to the free-text entry by maintenance technicians and a lack of constraints on spelling, grammar, or vocabulary. These difficulties are highlighted in (Devaney, Ram, Qiu, & Lee, 2005; Brundage et al., 2019). Maintenance personnel often prefer to enter data quickly, as their job depends on performing maintenance actions efficiently and correctly, rather than entering elaborate descriptions that adhere to formal language descriptions. Though methods have been proposed for identifying manufacturing issues from text (e.g., assembly issues (Madhusudanan, Gurumoorthy, & Chakrabarti, 2017)), this paper focuses on relations between natural language of MWOs and maintenance times.

Once properly processed and prepared, MWO data can be used to compare various possible maintenance actions in regards to expected outcome, duration, cost, or other resource investments. For example, Bokinsky et al. (2013) indicated the need to compare the actual performed maintenance action versus what was listed in a manual to ensure best practices are being upheld. This conclusion followed from a natural language analysis of Maintenance Action Forms for aircraft to reduce the time it is out of service. The timeliness of maintenance is also stressed in (Parida & Kumar, 2006), such as the role of downtime, that is directly related to the maintenance activity being undertaken. In the domain of medical devices, Sipos et al. (2014) built predictive models for equipment failures from billions of event logs.

A need exists to further analyze the rich yet under-explored knowledge of MWOs. In particular, there is a need to capitalize on the potential usefulness of capturing and understanding

Madhusudanan Navinchandran, Fnu; Sharp, Michael; Brundage, Michael; Sexton, Thurston.
"Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

time-durations of maintenance activities. This paper, as described in the next section, discusses two potential methods to analyze MWO datasets.

3. METHODOLOGY

This work coalesces free-form information found in MWOs into annotated concepts that are designated into the categories *problems*, *solutions* and *items* using Nestor, a previously developed tool (Sexton et al., 2017). In this context, *problems* are faults, failures, symptoms, or other motives inspiring the MWO. *Items* are the location or equipment where the *problem* is observed or work is performed, and *solutions* are the maintenance action taken. This work studies the *solutions* and *items* that are most important to time spent on maintenance activities. Firstly, it is necessary to clean and identify the various features of MWOs, such as the various *problem* features (e.g. *fault, leak*), *solution* features (e.g., *debur, reset*) and *item* features (e.g., *cylinder, flywheel*). Then, predictive models are built with these features as predictor variables. Based on the models, the most important features are then identified by studying how these features influence the overall maintenance time duration.

3.1. Text cleaning using tagging

A combination of human annotation and computer assistance in the form of Nestor is used to identify the various *problem*, *solution* and *item* features contained in the MWOs. The produced tags, or ideological concepts, are clean representations of noisy unstructured MWO text. For example, *Replace*, which is an alias for all related indications (e.g., Repalce <=> Replace <=> Replacing), is tagged as a *SOLUTION*. Use of tagged words lowers the variations in morphological forms and spellings as compared to raw MWO text, since a human has clarified their correct spelling with the correct alias. For example, in the dataset described in this paper (see Section 4.1), action words (verbs) extracted by Natural Language Processing resulted in 577 *solution* words, whereas tagging resulted in only 65.

3.2. Pipeline for Predictive Model

A pipeline of machine learning algorithms is used to build models for predicting maintenance durations using MWO knowledge. Instead of trying to predict exact times, explainable time classes were used as target variables. The use of time classes has two advantages. Firstly, it provides better performance than when trying to predict exact time values. Secondly, is more intuitive to understand and useful for maintenance managers since the classes give relatable quantities of task duration that provide reasonable expectations of accuracy.

Two separate variations of predictive models were investigated, neural networks and decision trees, and compared for ease of construction, flexibility, accuracy, and intuitive interpretation. Each model pipeline is structured to capitalize on the individual strengths of their respective base models.

3.2.1. Neural Network Model

The first step in the neural network driven pipeline is a set of neural networks called autoencoders. An autoencoder is '*a neural model where output units are directly connected with or identical to input units*' (Li, Luong, & Jurafsky, 2015). Autoencoders enable abstracting input features into more condensed and consistent representations. These were selected due to their ease of setup and ability to efficiently condense related concepts using only the most relevant data. Next a binary classifier is used to obtain a broad prediction of time classes. Finally, each broad class is processed using classifiers to get a finer time class for maintenance.

3.2.2. Decision Tree Model

A decision tree classifier was also investigated and compared to the neural network model. Decision trees were selected both for their intuitive structure in relating input importance, and their speed and strength for classification (this will be described in more detail in section 4.4). With the solutions, problems and items as features, and various time classes as outputs, a decision tree classifier would result in a model to predict any one of the time classes.

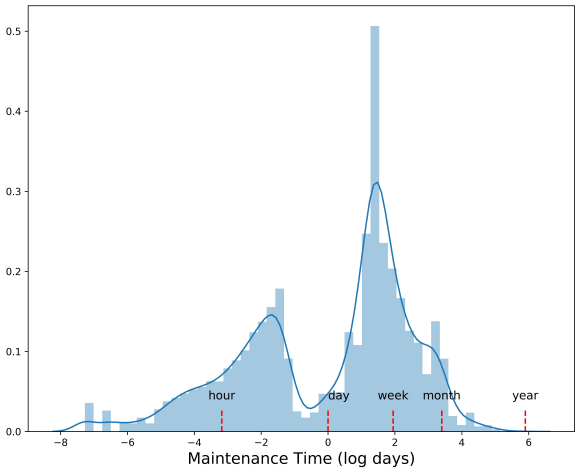
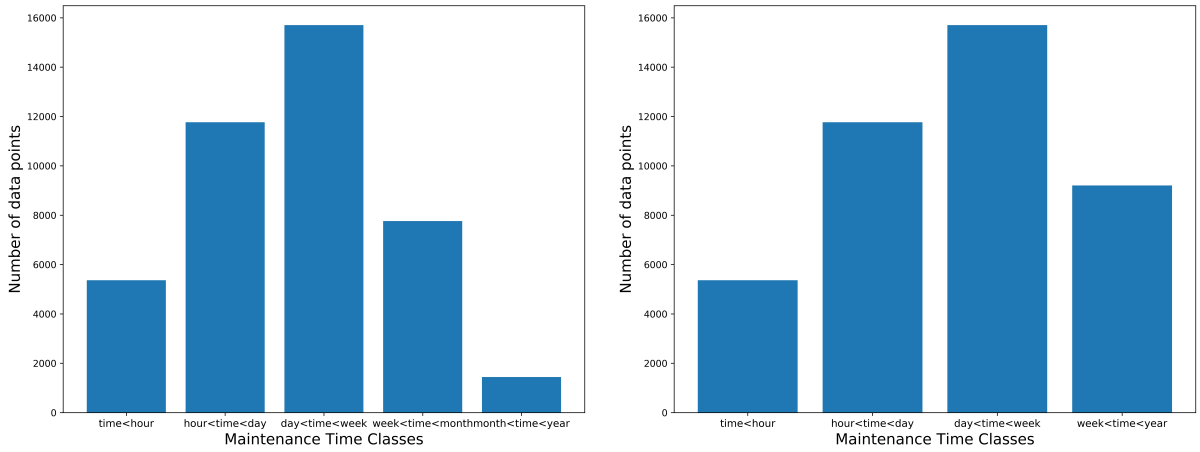


Figure 1. Distribution of times for the dataset. Various timescales are indicated using markers. (The y-axis represents the kernel density estimate for the time values).

3.3. Analysis and Model Interpretation

Once classification models are built for the maintenance features, the structure and behavior of the models themselves be-

Madhusudanan Navinchandran, Fnu; Sharp, Michael; Brundage, Michael; Sexton, Thurston.
"Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.



(a) Five Time Classes (Average Count = 8409)

(b) Four Time Classes (Average Count = 10511)

Figure 2. Distribution of workorder samples across different time classes. The distribution varies widely with some classes having low number of samples, which are balanced after resampling to the average number of samples.

come the subject of analysis and can be used to determine the most influential input features found in the original MWOs. In order to help determine the most important features, a sensitivity analysis was performed by monitoring the individual models performance when each feature is excluded during training, one at a time. This method is used since it reveals the most influential features on model performance during the construction phase. This is not the only indicator of importance to the relationship between maintenance action duration and the captured concept features but it is a good estimation to help guide further investigations. There are some other methods for sensitivity and importance analysis that are useful - these are discussed in Section 5. For the decision tree, additional information regarding the most influential features is found by looking at the importance of features in the decision tree structure itself.

With this generic methodology in place, we now describe a case study that illustrates the application of the methodology to analyze a real manufacturing dataset.

4. PREDICTIVE MODELS - A CASE STUDY

This section illustrates the application of predictive models described in the previous section on a manufacturing dataset and presents the performances of the models.

4.1. Dataset used

The MWO dataset used for this study was sourced from a real automotive manufacturer and consisted of 47 798 MWOs. The dataset is in a spreadsheet format, and some of the fields are *Workorder Number*, *Status*, *Actual Start Date and Time*, *Actual Finish Date and Time*, *Asset Number*, Textual *Description*, *Location* and *Reported By*. More details about information contained in MWOs can be found in (Brundage, Morris, Sexton, Moccozet, & Hoffman, 2018). The most important fields from these MWOs are the actual start and finish times and the two text description fields about the maintenance activity.

4.1.1. Data Quality Challenges

In preparation for the analysis portion of this work, the non-uniformity of the dataset presented some unique challenges that are likely to be common in real world applications. Every MWO dataset has its own characteristic fields, such as *Asset Number, Workorder Number, Problem Description, Requested By, Solved By, Opening Time* and *Cost Incurred*. For this research, the text descriptions of the problems and actions taken and time-related fields are of interest. The text descriptions were annotated with tags using the methods described in (Sexton et al., 2017).

MWOs contain dates and times in different formats and hence must be preprocessed to get maintenance time durations. To improve consistency, all time data is converted to days with a range across the data set spanning from zero to hundreds of days. The dataset used for this paper had only the starting and ending times for the workorders, not the actual task work hours. Hence, it was not possible to ascertain the actual duration of maintenance solutions. The assumed duration is deemed to be the end time minus the start time listed on the MWO. These time calculations do not always accurately reflect the actual time taken for maintenance, but are instead rough estimates due to inconsistencies and variations

Madhusudanan Navinchandran, Fnu; Sharp, Michael; Brundage, Michael; Sexton, Thurston.
"Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

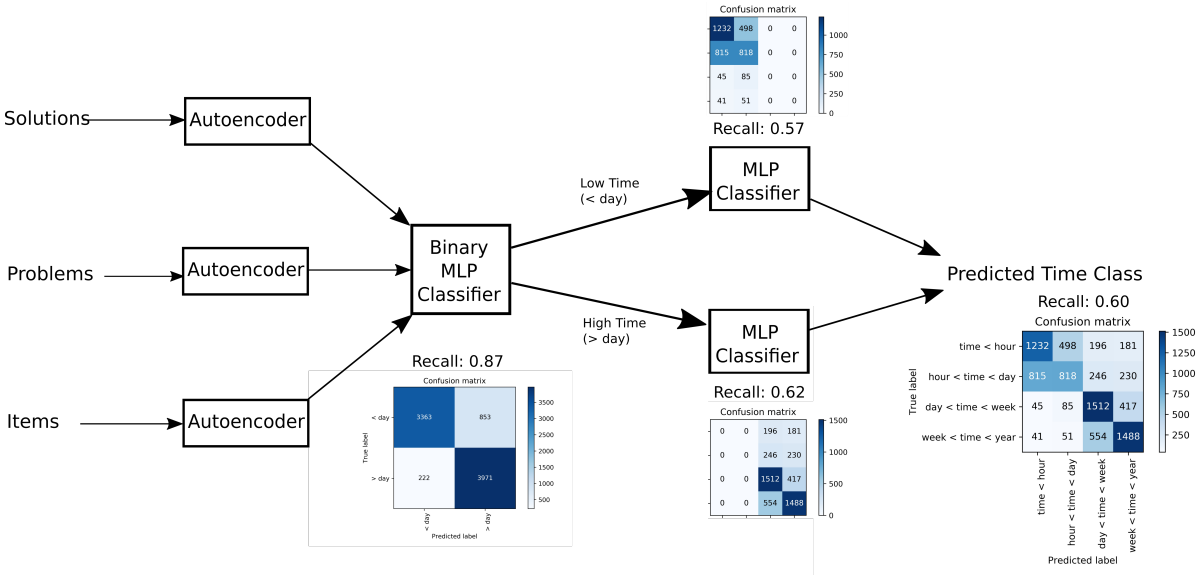


Figure 3. The schematic of the architecture of the classifier model used. Recall scores at each classification step, along with the confusion matrices are also shown.

in recording times. Often, missing time entries exist for some maintenance activities. For this case study, there are 4 914 entries with incorrect formats and missing time entries. There are a further 843 entries with negative total duration. These are excluded for the purpose of analysis. More fine-tuned, and correct time recordings are needed to improve the accuracy of these resulting models. More details about this aspect are presented in Section 5.

4.2. Time distributions and time classes

The intended output for the predictive models is a categorical window of the maintenance duration. Though maintenance times in MWOs are real-valued, predicting the output times as precise real-valued numbers is far less useful than practical task assignment windows because jobs are typically scheduled into some window of an expected duration of the task. In other words, short tasks may be scheduled in 5 min blocks, but longer tasks are more commonly blocked off in terms of hours. For this data set, some concessions of the designation of the duration categories is also fed from the small number of data points and low accuracy of predictions found in initial studies with the time data. With larger volumes of more accurate data this could be overcome. Despite these concessions, the conclusions and methods developed in this work could easily be extended to the level of granularity most useful and feasible for any target use case.

The designation of the classes in this work was largely led by expert intuition and observations of the data itself. The distribution of times shown in Figure 1 makes it clear that there is a large split in amount of actions requiring less than a day, and another smaller cluster break between the week and the month markers. Additionally, practical and relatable demarcations such as *hours*, *days*, *weeks*, *months* and *years* are more explainable and help foster understandings of the ensuing analysis than simply putting across numerical ranges of times.

Figure 2(a) shows the distribution of frequencies of samples across five classes (*hour, day, week, month,* and *year*). The data points were resampled across all classes to match the average number of samples across all classes. Since there are a relatively small number of values in the fifth class (month<time<year), it can also be combined into the fourth class and represented as a single time class (week<time<year). Figure 2(b) shows the same distributions if there are only four classes, corresponding to *hour, day, week* and *year*. The four classes consist of:

- Less than an hour ($<$ 1 h)
- Greater than an hour but less than a day (1 - 24 h)
- Greater than a day but less than a week (24 - 168 h)
- Greater than a week ($>$ 168 h).

4.3. Application of pipelines to manufacturing dataset

With the *solution, problem* and *item* tags as features and time classes as outputs, classifiers are built to predict the time classes for maintenance. A schematic for the entire pipeline is shown in Figure 3. The first step is to preprocess the input features using a set of autoencoders. These autoencoders

Madhusudanan Navinchandran, Fnu; Sharp, Michael; Brundage, Michael; Sexton, Thurston.
"Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

are used to compress these features into approximately half the original number of inputs. The autoencoders are used to both focus the classifiers and help to remove tangential information contained in the original feature set. The number of *problems* reduces from 51 to 32, *solutions* from 65 to 32, and *items* from 271 to 128.

The first stage binary classifier was set to delineate at the largest gap in the original data distribution, the less than one day mark. The purpose of this classifier was to roughly judge if the jobs were 'short' or 'long' based on the language found in the MWO. The classifier implemented is a Multi-Layer Perceptron (MLP) with three hidden layers. MLP is a neural network with an input layer, an output layer and at least one hidden layer. The performance of a classifier is judged using the *recall* score, which is the fraction of all correct time classes correctly identified by the classifier. It calculated as the ratio of the number of *true positives* to the sum of *true positives* and *false negatives* (with an averaging method for multiclass classification). The split between training and testing sets was 80 % to 20 % respectively. The binary classifier performs with a recall score of 0.87 (For five classes, recall=0.88).

In the next step, each of the long time and short time classes are separately classified using two separate additional MLP classifiers. These additional classifiers respectively further classify each MWO input into the *hour* vs. *day* categories if the MWO was found to be a 'short' job or the *week* vs. *year* categories if it was designated a 'long' job.

The 'long job' (high time) MLP classifier (recall = 0.62) was better during testing than the 'short job' (low time) MLP classifier (recall = 0.57). When the predictions for all the classes are combined, the overall recall is 0.60 (For five classes, overall recall = 0.59).

4.4. Decision Tree Classification

The design and implementation of the previous pipeline of classifiers inspired a separate step-by-step classification approach, a decision tree classifier. Similar to the previous pipeline, the target output for the decision tree is the time classification of each MWO based on its captured language. Decision trees were built both with and without the use of auto encoders as preprocessing nodes to compare if there was any trade off between end analysis results vs model accuracy. The performance of decision tree classifiers (using only input features) is shown in Figure 4. The time classes corresponding to within an hour, week and year are more correctly classified than the in-between classes (day and month). In general, the decision tree classifier had better performance than MLP classifiers, with recall scores of

- 0.66 (only input features and 4 time classes)
- 0.67 (using autoencoders and 4 time classes)
- 0.64 (only input features and 5 time classes)
- 0.65 (using autoencoders and 5 time classes)

The use of autoencoders to preprocess the features improves the performance of the classifier slightly, but also adds a layer of obfuscation to the final sensitivity analysis that may outweigh the accuracy gain in many cases.

4.5. Most important features

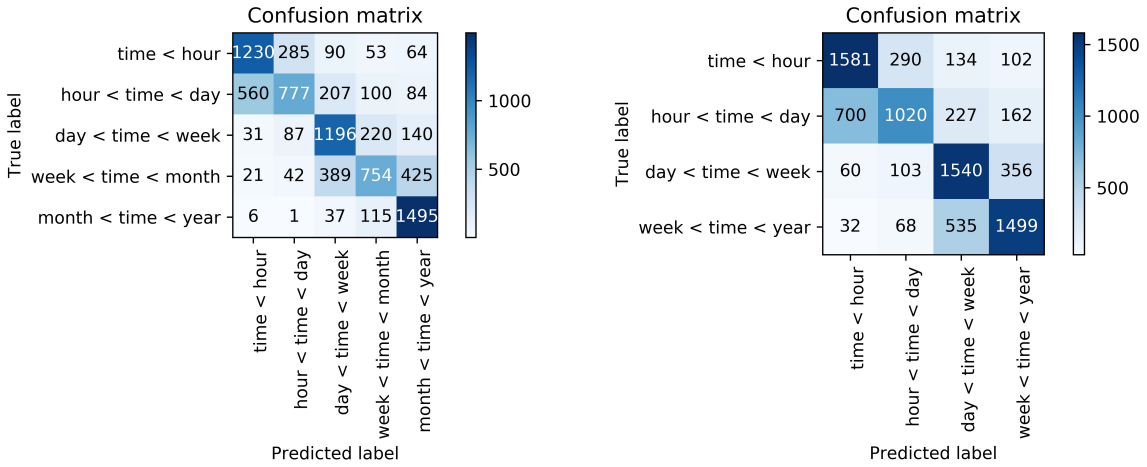
The models described earlier have targeted the prediction of estimated time for a maintenance activity. However, it is also useful to inform the maintenance manager about the most important features that influence the amount of time taken. There are two methods described here to decide the most important features. These methods are demonstrated on the decision tree classifier method described in the previous subsection.

In the first method, the decision tree classifier (with autoencoders for preprocessing) is supplied with all features, except one. This is repeated for all features, and the performance of the classifier is recorded. Figure 5 shows how the recall varies for each feature removed, for 387 features. The important features can be inferred from low recall values when the feature is absent.

As an alternate method, the important features for the decision tree classifier (without autoencoders) are obtained from their Gini importance (Shouman, Turner, & Stocker, 2011). This method did not use autoencoders since the output features from autoencoder are not uniquely identifiable. The various *problems*, *solutions* and *items* are shown in Figure 6(a). Since there are more items due to it being the largest category, the features are shown again in Figure 6(b) by dividing with the number of features in that category.

Between the two different methods there are some common features that are shown as important. Common solutions were *completed* and *clean*; some common problems are *fault* and *dirty* and some items are *beam, conveyor, hmi, primary, linebore* and *clamp*. This list is useful to a maintenance manager to help identify which solutions have the greatest influence on maintenance time. It is also useful to identify items that are anomalous with respect to time i.e. the maintenance for these items takes too long or are unusually quicker than expected. Deeper analysis for determining the root cause may then be undertaken. For example, the words *dirty* and *clean* have large importance - this could mean that depending on whether something was dirty and cleaned would strongly determine the amount of time for maintenance. Common sense tells us that cleaning is not a very time consuming activity as compared to, say, replacing an entire part. Similarly, items such as *conveyor* and *linebore* are consistently found to influence time durations. Such sensitivity analysis can be coupled with other models such as regression models to determine the

Madhusudanan Navinchandran, Fnu; Sharp, Michael; Brundage, Michael; Sexton, Thurston.
"Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.



(a) Five Time Classes

(b) Four Time Classes

Figure 4. Performance of Decision Tree Classifiers for five and four time classes.

nature of influence of important features - whether they contribute to increased (or reduced) time durations. This analysis could eventually lead to better planning and resource management by identifying and quantifying reasonable expectations regarding various maintenance tasks within a facility.

5. DISCUSSION AND FUTURE WORK

This paper has discussed the use of machine learning models to predict time classes for maintenance duration. Numerous factors influenced the performance and results of the feature importance analysis. Some of the notable observations regarding that are listed here.

Assignment of Duration Categories: During the course of the investigation, various demarcations of time duration categories were explored to best describe the data. For the MLP classifier, the binary classifier performance decreased by reducing number of classes from 5 to 4. For the overall classifier, as well as decision tree classifier, reducing the number of classes led to improved performance. Also, within the multi-step classification of the MLP classifier, the performance is noticeably better for the binary classifier than for the second classification step for four/five classes. These specific results are somewhat dataset dependant, but the general trend of searching for classes that are adjacent to each other are largely expected to to improve performance regardless of the dataset.

Use of Tagged Data: The analysis illustrated the value of using tagged data, since the number of features to be used reduces significantly. Apart from clarifying the terms used, tagged data makes the analysis computationally feasible in terms of having to use lesser features. It also makes the results more explainable and coherent.

Quality of Time Data: Unusable data entries are a major issue. These are either missing time entries or incorrect entries, such as closing times for workorders that are earlier than opening times. MWOs for which there were missing dates and times were entirely ignored but this reduces the amount of correct data available to train the models. Such issues emphasize the importance of collecting accurate time related data during maintenance.

Maintenance Data Collection: The time data available in the dataset were only the actual start and finish times. There is no more specific time information e.g., when the maintenance technician arrived, or when the workorder was opened. Such finer time data would lead to improved inferences about the actual duration of maintenance. Further details of what other time data are useful can be found in (Brundage et al., 2018). Efficient data collection strategies are needed for better maintenance time data capture.

5.1. Application to maintenance management

The general procedure for analysis is an outcome of this work. To derive maintenance management insights starting from workorder data, the following procedure is a recommended workflow:

Clean the data by using Nestor to tag the data. This preprocessing results in a representation of data that is possible to be analyzed.

Calculate number of time classes for maintenance time du-

Madhusudanan Navinchandran, Fnu; Sharp, Michael; Brundage, Michael; Sexton, Thurston.
"Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

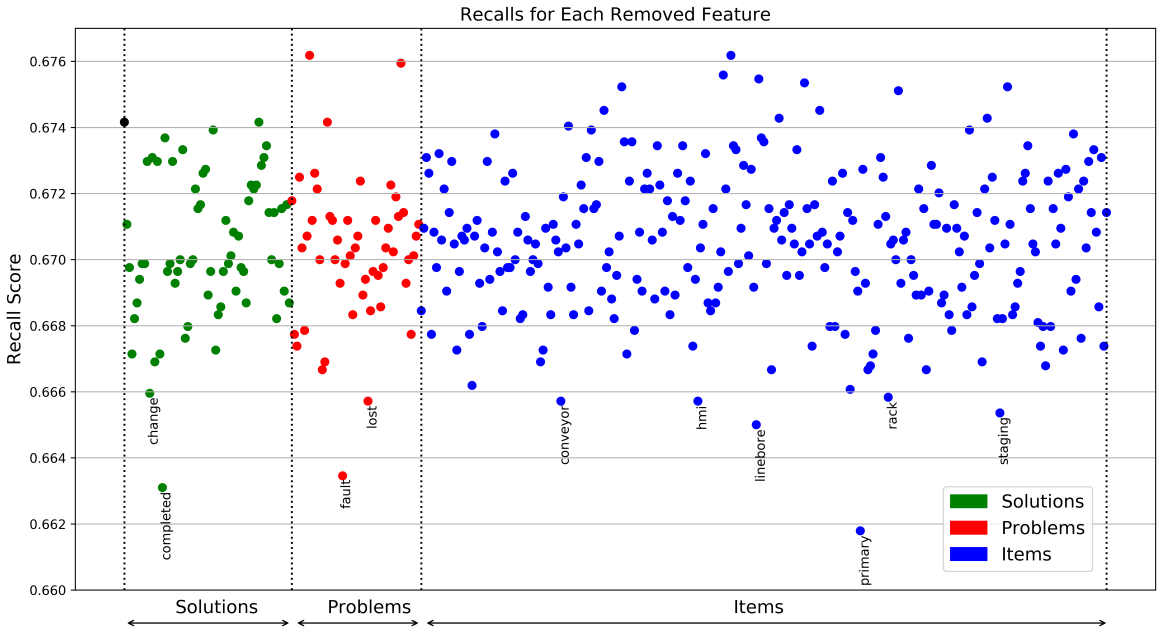


Figure 5. Performance of the decision tree classifier (with autoencoders for preprocessing) by removing one feature at a time. Some of the low recalls are labelled with the feature that is removed at that point.

rations. An appropriate number is chosen by looking at the distribution of times such that the number of entries in each class is comparable.

Choose a machine learning classifier depending on computing resources, dataset size and expected performance levels. A couple of examples have been discussed here, but there are many more options available. The use of machine learning models also involves splitting data for training and testing (For this paper the split was 80 % to 20 % respectively).

List out important features using methods such as feature importances for decision tree classifier. This would help maintenance managers identify hotspots such as certain items that have been contributing heavily to maintenance duration, even when it is not apparent without the analysis.

Identification of important features such as *problem* features, *solution* features or *item* features will help to address specific points in maintenance.

5.2. Future work

The machine learning models can be improved and more generalized observations can be derived. This is possible with larger and varied datasets and more tagging on datasets (such as greater time spent on tagging and tagging of bigram phrases). It could lead to generalized observations of tags that are indicative of maintenance domain. For example, some terms might relate to expensive or time consuming solutions, such as needed a *replacement* part. This could potentially contribute to language standards for MWO recording practices in maintenance.

With regards to the sensitivity of the models to input features, it is also planned to use the method of leaving out one feature at a time on the Neural Network Model. Also, there are other methods that could be used, for example, one could change the values of only one feature at a time to know the effect on predictions. Another method is to use partial dependence plots for visualizing importance of a given feature, one or two at a time. Use of such multiple methods might help to generalize the list of important features. These will all be addressed in future work.

This work utilized a dataset from the domain of manufacturing maintenance. Similar analyses can be performed on MWOs from other domains, such as aerospace, shipping, heating ventilation and cooling (HVAC), to identify common and domain specific parts of language that relate to the maintenance duration.

Wherever data is available, similar models can be built and studied for cost of maintenance activity.

5.3. Other applications

Beyond identifying important features, predictions of maintenance time could be useful for maintenance and production

Madhusudanan Navinchandran, Fnu; Sharp, Michael; Brundage, Michael; Sexton, Thurston.
"Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

scheduling. Time windows of maintenance obtained from the model can be used as input to decide how long a resource may be unavailable.

Based on a problem condition at a machine, a maintenance manager could search through MWOs for previous solutions. However, there could be many solutions, and no way to prioritize which solution to perform. The solutions to a problem may be about merely lubricating a part, or entirely replacing a component. Modeling the relation between the solutions, problems and items provides a means of ordering these solutions by the amount of time to be likely consumed. Prioritizing might lead to potential savings in overall time.

There are potential uses from this work with respect to managing the inventory of spares. If there are items that need frequent replacement which influence maintenance times, these items could be stocked in spares to shorten the duration.

The distributions of solutions, problems, items and times could be used to build simulation models of higher fidelity that have multiple failure modes and treatments. Thus, simulations of maintenance scheduling might be more well informed.

6. CONCLUSIONS

This paper discussed the analysis of MWO data to help estimate maintenance duration and identify important *problems*, *solutions* and *items* features. These features are used to build machine learning models to predict estimated time duration of maintenance activities. From the machine learning models, it is also possible to infer important features that influence maintenance time. The methodology for using MWO data to infer important features involves cleaning the text data, deciding on the time classes, fitting predictive models and listing most important features from the models.

ACKNOWLEDGMENT

The authors wish to acknowledge the inputs and support of Sascha Moccozet in the initial phases of this work.

DISCLAIMER

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

REFERENCES

Ahmad, R., & Kamaruddin, S. (2012). An overview of time-based and condition-based maintenance in industrial application. *Computers & Industrial Engineering*, *63*(1), 135–149.

Allen, R. (2013). Tribal knowledge. *Quality*, *52*(1), 54–59.

Bokinsky, H., McKenzie, A., Bayoumi, A., McCaslin, R., Patterson, A., Matthews, M., ... Eisner, L. (2013). Application of natural language processing techniques to marine v-22 maintenance data for populating a cbm-oriented database. In *Ahs airworthiness, cbm, and hums specialists' meeting, huntsville, al.*

Brundage, M. P., Morris, K., Sexton, T., Moccozet, S., & Hoffman, M. (2018). Developing maintenance key performance indicators from maintenance work order data. In *Asme 2018 13th international manufacturing science and engineering conference* (pp. V003T02A027–V003T02A027).

Brundage, M. P., Sexton, T., Hodkiewicz, M., Morris, K., Arinez, J., Ameri, F., ... Xiao, G. (2019). Where do we start? guidance for technology implementation in maintenance management for manufacturing. *Journal of Manufacturing Science and Engineering*, 1–24.

Camci, F. (2015). Maintenance scheduling of geographically distributed assets with prognostics information. *European Journal of Operational Research*, *245*(2), 506–516.

Devaney, M., Ram, A., Qiu, H., & Lee, J. (2005). Preventing failures by mining maintenance logs with case-based reasoning. In *Proceedings of the 59th meeting of the society for machinery failure prevention technology (mfpt-59).*

Gulati, R., & Smith, R. (2009). *Maintenance and reliability best practices*. Industrial Press Inc.

Helu, M., & Weiss, B. (2016). The current state of sensing, health management, and control for small-to-medium-sized manufacturers. In *Asme 2016 11th international manufacturing science and engineering conference* (pp. V002T04A007–V002T04A007).

Li, J., Luong, T., & Jurafsky, D. (2015). A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (Vol. 1, pp. 1106–1115).

Locks, M. O. (1973). *Reliability, maintainability and availability assessment*. Hayden.

Madhusudanan, N., Gurumoorthy, B., & Chakrabarti, A. (2017). Automatic expert knowledge acquisition from text for closing the knowledge loop in plm. *International Journal of Product Lifecycle Management (IJPLM)*, *10*(4).

Mukherjee, S., & Chakraborty, A. (2007). Automated fault tree generation: bridging reliability with text mining. In *Reliability and maintainability symposium, 2007. rams'07. annual* (pp. 83–88).

Parida, A., & Kumar, U. (2006). Maintenance performance measurement (mpm): issues and challenges. *Journal of Quality in Maintenance Engineering*, *12*(3), 239–251.

Sexton, T., Brundage, M. P., Hoffman, M., & Morris, K. C.

Madhusudanan Navinchandran, Fnu; Sharp, Michael; Brundage, Michael; Sexton, Thurston.
"Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

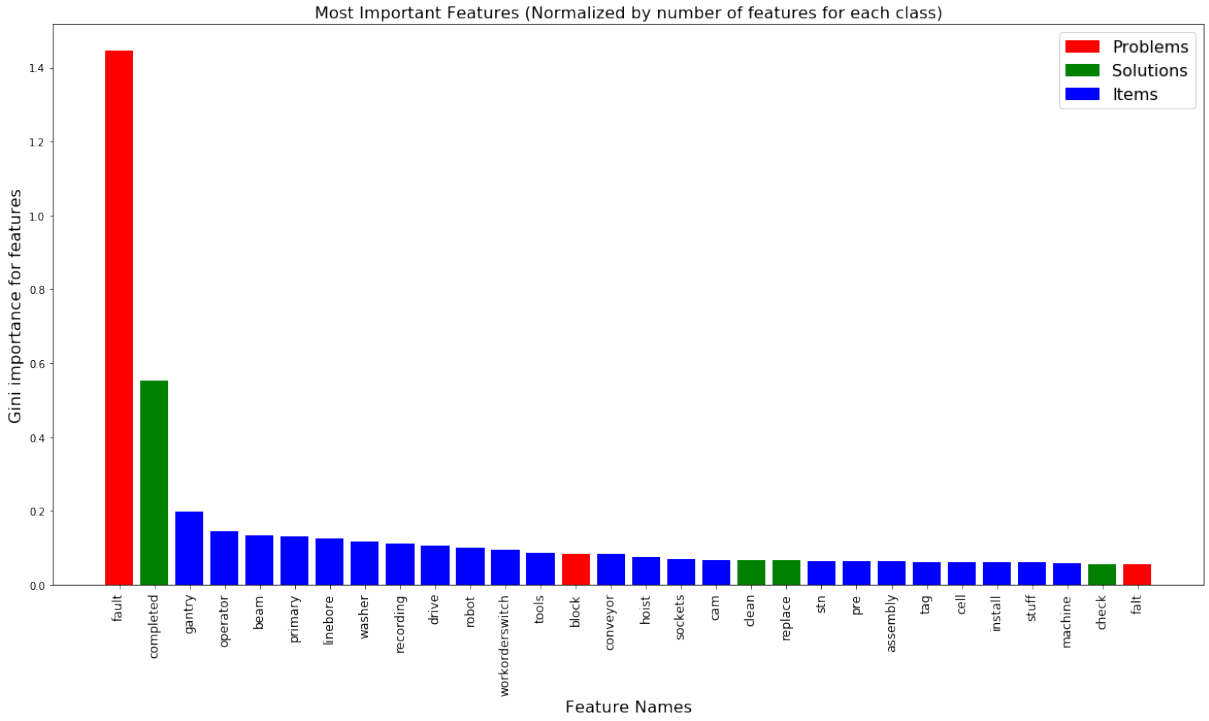
(2017). Hybrid datafication of maintenance logs from ai-assisted human tags. In *Big data (big data), 2017 ieee international conference on* (pp. 1769–1777).

Sherwin, D. (2000). A review of overall models for maintenance management. *Journal of quality in maintenance engineering*, *6*(3), 138–164.

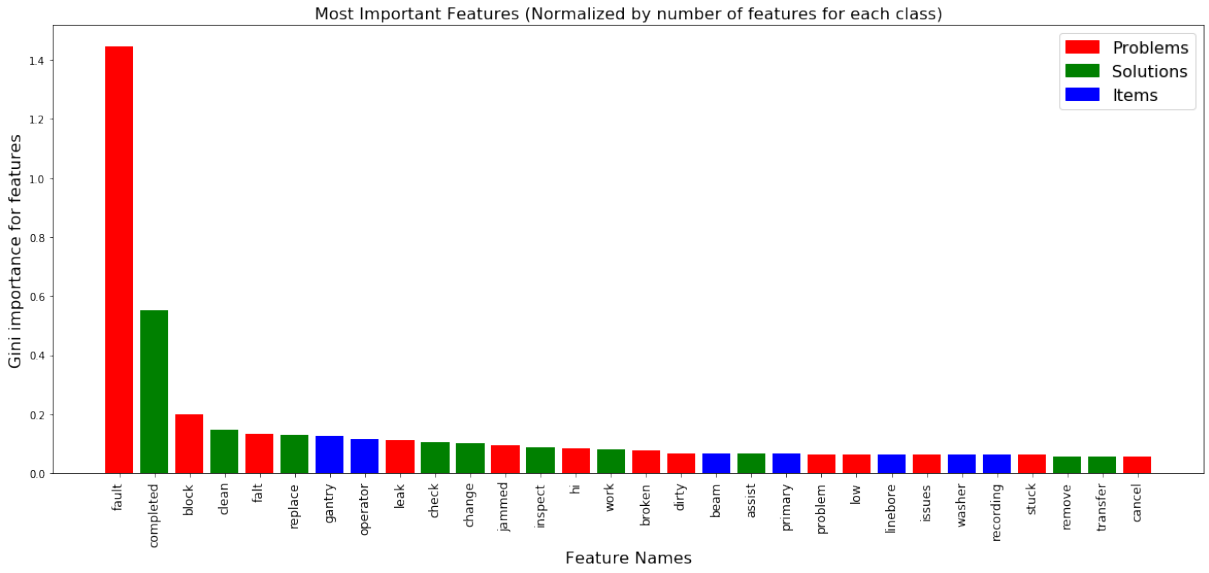
Shouman, M., Turner, T., & Stocker, R. (2011). Using decision tree for diagnosing heart disease patients. In *Proceedings of the ninth australasian data mining conference-volume 121* (pp. 23–30).

Sipos, R., Fradkin, D., Moerchen, F., & Wang, Z. (2014). Log-based predictive maintenance. In *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1867–1876).

Madhusudanan Navinchandran, Fnu; Sharp, Michael; Brundage, Michael; Sexton, Thurston.
"Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.



(a) (without normalizing).



(b) (after normalizing using number of features).

Figure 6. Ordered list of most important features from the decision tree classifier. These features are ranked by their order of Gini importance. In Figure (a), the top 30 features are simply ranked by importance. Since the number of features is maximum for *items*, most of the important features are *items*. Hence, Figure (b) shows an ordered list, where the importance is divided by the number of either *problems*, *solutions* or *items*.

Madhusudanan Navinchandran, Fnu; Sharp, Michael; Brundage, Michael; Sexton, Thurston.
"Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining

Emily M. Hastings[1], Thurston Sexton[2], Michael P. Brundage[3], and Melinda Hodkiewicz[4]

[1] *University of Illinois, Urbana, IL 61801, USA*
ehstngs2@illinois.edu

[2,3] *National Institute of Standards and Technology, Gaithersburg, MD 20814, USA*
thurston.sexton@nist.gov
michael.brundage@nist.gov

[4] *University of Western Australia, Crawley WA 6009, AUS*
melinda.hodkiewicz@uwa.edu.au

ABSTRACT

Maintenance work orders (MWOs) are an integral part of the maintenance workflow. These documents allow technicians to capture vital aspects of a maintenance job, including observed symptoms, potential causes, and solutions implemented. MWOs have often been disregarded during analysis because of the unstructured nature of the text they contain. However, research efforts have recently emerged that clean MWOs for analysis. One such approach is a tagging method which relies on experts classifying and annotating the words used in the MWOs. This method greatly reduces the volume of words used in the MWOs and links words, including misspellings, that have the same or similar meanings. However, one issue with this approach and with the practical usage of data-annotation tools on the shop-floor more generally is the usage of only one expert annotator at a time. How do we know that the classifications of a single annotator are correct, or if it is, for example, feasible to divide the tagging task among multiple experts? This paper examines the agreement behavior of multiple isolated experts classifying and annotating MWO data, and provides implications for implementing this tagging technique in authentic contexts. The results described here will help improve MWO classification leading to more accurate analysis of MWOs for decision-making support.

1. INTRODUCTION

Maintenance Work Orders (MWOs) are one of the main records of activities that occur during a maintenance event. MWOs often include data such as problems observed, corrective actions taken, potential causes, necessary parts, and time of machine breakdown. The information in these MWOs is useful in maintenance decision making; for example, it can be used in failure mode identification, problem spot identification, and calculating more accurate mean time to repair (MTTR) and mean time between failure (MTBF) metrics, which can improve maintenance strategy and efficiency (Sexton, Hodkiewicz, Brundage, & Smoker, 2018). However, the MWO data, in its raw form, is often too unstructured, informal, and filled with jargon for immediate analysis. To combat this issue, researchers have used Natural Language Processing (NLP) techniques to extract important information out of the natural language descriptions within the MWOs. Examples of recent work in this area include (Smoker, French, Liu, & Hodkiewicz, 2017; Sexton, Brundage, Morris, & Hoffman, 2017; Brundage, Morris, Sexton, Moccozet, & Hoffman, 2018; Sexton et al., 2018; Brundage, Kulvatunyou, Ademujimi, & Rakshith, 2017).

One promising area in this space involves "tagging" the MWOs, by assigning "tags" to concepts of importance in the maintenance domain (Sexton et al., 2018). For example, a MWO might contain the text "Replaced the hydraulic hose and fixed the leaking valve." The important concepts to a maintenance practitioner might be the problem that was observed, the items that were addressed, and the solutions that were provided. In this example, the problem would consist of "leak", the items are "hydraulic hose" and "valve", and the solutions are "replaced" and "fixed". Researchers at the National Institute of Standards and Technology (NIST) have created an open-source, free toolkit called Nestor[1] to aid in

[1] https://github.com/usnistgov/nestor

Hastings, Emily; Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda.
"Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

this tagging process by estimating the a priori importance of concepts in the corpus, and helping annotators link together potential cases of domain-specific abbreviations, misspellings, and synonyms. Once tagged, this data can then be used to apply the previously-inaccessible knowledge stored in the MWOs to improve the manufacturing process by, for example, diagnosing and addressing problems faster (Sexton et al., 2017).

The current research on this tagging approach to MWO analysis focuses primarily on introducing the method and on issues of the quality of the MWOs themselves, and has not yet examined more practical issues related to the deployment of the approach in authentic maintenance contexts. For example, this research assumes that a single person will annotate a given dataset through tagging. Indeed, it is likely the case that a single person would be assigned this task in an authentic maintenance environment that uses this technique. However, this assumption raises questions that need to be addressed. For example, is a single isolated tagger reliable or experienced enough to properly tag the dataset and produce data usable for future analysis? In addition, it may not always be the case during the use of a tagging tool that a single individual tags the entire dataset. This tagger may have only limited time to devote to the task, and so it may be necessary for additional taggers to contribute their own work later. In this alternate situation, can multiple taggers achieve sufficient agreement with each other to produce usable data and warrant splitting the task? Since the information gained from analyzing the tagged MWO data will be used in important maintenance decisions, it is crucial that these tags be accurate and reliable. In the absence of a gold standard for tagging, it is necessary to utilize an alternative validation method to answer these questions.

This paper seeks to address these issues by investigating the agreement behavior of multiple isolated experts tagging MWO data. We conducted an experiment in which six annotators independently tagged a single dataset using the Nestor toolkit. By having multiple people tag the data, agreement on classifications of the different words in MWOs (e.g., "replace" is a solution) and the aliases assigned to them (e.g., "replace," "replaced," and "repalce" refer to the same concept) could be measured to assess the level of consensus reached and the viability of using a tagging approach to analysing MWO data in practice. We found that the six annotators achieved high levels of agreement, lending support to the use of a tagging approach to clean MWO data for analysis. We also identified several opportunities for improvement of the approach; for example, performing the tagging task in an environment that supports real-time feedback or collaboration could further improve the level of consensus achieved.

1.1. Background on Crowdsourcing

These issues of agreement among multiple taggers are common in the domain of crowdsourcing, where complex jobs are broken down into smaller, granular tasks and completed individually by many people, whose work is aggregated to create a final solution (Surowiecki, 2005). A common application of crowdsourcing is generating labeled training data to be used for machine learning algorithms, for example, labelling the contents of images (Nowak & Rüger, 2010) or the emotion of speech assets (Tarasov, Delany, & Cullen, 2010). Crowdsourcing has been shown to be an effective method for generating high-quality labels cheaply and efficiently (Hsueh, Melville, & Sindhwani, 2009; Snow, O'Connor, Jurafsky, & Ng, 2008; Ambati, Vogel, & Carbonell, 2010).

The concept of inter-annotator agreement or reliability is crucial to this line of work, as it can give researchers a clearer idea of whether multiple annotations are needed for each piece of data, whether non-expert annotators are capable of providing reasonable labels, and similar knowledge (Nowak & Rüger, 2010; Brew, Greene, & Cunningham, 2010). Some common metrics of agreement used in this context are accuracy (e.g., (Nowak & Rüger, 2010; McCreadie, Macdonald, & Ounis, 2010)), the κ-statistic (Fleiss, 1971; Randolph, 2005), and correlation coefficients such as Kendall's τ (Kendall, 1938).

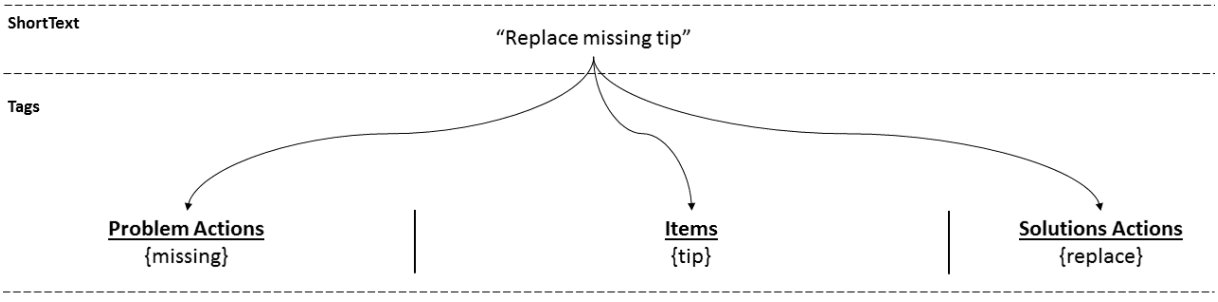


Figure 1. Taken from (Sexton et al., 2018), to illustrate the procedure used both in that work and in the case study presented here, for tagging a MWO with Nestor. Tokens are extracted from the original MWOs, and annotators are tasked with mapping each to an alias and a classification, which together form a "tag".

Hastings, Emily; Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda.
"Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

We ground our work in this body of prior literature and present a study reporting the levels of agreement reached by multiple experts tagging MWO data. We also provide insights into how to best measure agreement in this context.

1.2. The Tagging Tool

As described in (Sexton et al., 2018), the Nestor toolkit, used here to characterize machine-assisted tagging of MWOs, requires input of existing data to be processed, along with an assumed tag schema that represents the possible types of tag classifications within the corpus of MWOs. Here we utilize the toolkit default schema—namely, *Item*, *Problem*, and *Solution* tags, along with *Unknown* tags where some ambiguity exists without more context.

The mapping task performed by an annotator within Nestor thus consists of taking a list of extracted "tokens" — the word-level strings of text that were estimated to be statistically important per the Nestor back-end — and giving them an alias and a classification (see Figure 1). Importance in this case is given by the tokens' term-frequency/inverse-document frequency (TF-IDF) score, a common NLP metric (Leskovec, Rajaraman, & Ullman, 2014). Tokens are presented in decreasing order of importance to facilitate the efficiency of the annotator's task, and potentially related tokens that likely share an alias are recommended by a fuzzy string match. The final output is then a sort of "dictionary" that can parse out useful tags from a large variety of more informal, jargon-filled or misspelled text documents.

2. EXPERIMENTAL DESIGN

To investigate the agreement behavior of multiple users tagging MWO data, we conducted a study with 6 expert annotators from NIST and University of Western Australia. Using the "Research Mode" of the Nestor toolkit, participants tagged MWO data from a publicly available dataset about excavation machinery, which is included with the tool[2]. Research mode was used because it allows the tool to periodically and automatically save the classifications and aliases assigned so far, in order to provide insight into trends in agreement over time.

Participant volunteers were to tag the dataset using the tool's "Single-word Analysis" section for approximately 30 minutes. This task length was chosen in order to allow participants to tag a sufficient amount of data for analysis while reducing the risk of performance loss due to fatigue. Prior work has shown that factors such as vigilance decrement can lead to reduced performance during long tasks after approximately 30 minutes, especially when the tasks require discrimination based on a standard held in memory, as the tagging does (Mackworth et al., 1950; Parasuraman & Davies, 1976; Parasuraman, 1979). The task was also limited to this length in order to minimize intra-study training effects, which can occur as participants become better at experimental tasks over time. Due to the design of the tool, the highest rate of annotation occurs at the beginning of the task (with the most important tokens) (Sexton et al., 2018), and so a large amount of tagging can be completed while training effects are at their lowest.

[2]The dataset can be found at https://prognosticsdl.ecm.uwa.edu.au/pdl/ labeled *ExcavatorBucketFailures*. If using this data, please provide the proper citation.

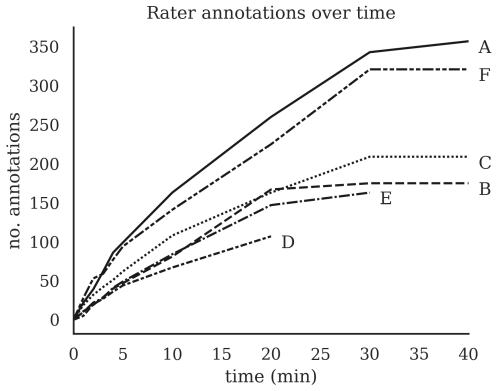


Figure 2. Participants' annotations over time.

3. RESULTS

On average, participants each completed 265 annotations during the allotted time, with a low of 175 and a high of 357. Figure 2 shows participants' progress over time. Interestingly, although participants had been instructed to perform the tagging task for 30 minutes, there was some variability in the actual amount of time spent tagging. We also observed differences in tagging rate between participants, as might be expected.

3.1. Agreement Measure

To measure the level of agreement achieved by taggers we use Fleiss' Multi-rater Kappa statistic (κ_{Fleiss}) (Fleiss, 1971). κ_{Fleiss} is given by

$$\kappa_f = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}\,, \tag{1}$$

where $\bar{P}$ represents the proportion of overall observed agreement, and $\bar{P}_e$ represents the proportion of agreement between raters expected by chance.

$\bar{P}$ is given by

$$\bar{P} = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij^2} - Nn\right), \tag{2}$$

Hastings, Emily; Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda.
"Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

and $\bar{P}_e$ is given by

$$\bar{P}_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \right)^2 , \quad (3)$$

where N is the number of cases, n is the number of ratings per case, and k is the number of rating categories. It is important to note that when calculating κ_{Fleiss}, not all raters need to provide a rating for every case.

Fleiss suggests that values of κ_{Fleiss} less than 0.4 indicate low agreement, values between 0.4 and 0.7 indicate good agreement, and values over 0.7 indicate excellent agreement, although these conventions vary. Other research suggests that only values of κ_{Fleiss} above roughly 0.7 indicate good agreement (Tarasov et al., 2010).

3.2. Alias Agreement

We first consider the raters' agreement on the alias assigned to each token. Participants achieved a κ_{Fleiss} of 0.85, which indicates a high level of agreement.

In addition to this analysis of the entire set of aliases produced, we also calculated the agreement on only the most important tags. These aliases were those that were associated with tokens the tool ranked in top 1 % in terms of statistical importance (recall that the tool ranks tokens' importance via TF-IDF scores). For this set, κ_{Fleiss} was 0.81, which also indicates good agreement.

We further analyzed changes in the level of agreement over time, as shown in the top section of Figure 3. The figure shows these trends for both the full set of tokens and the reduced set of important tokens. See the left side of Figure 4 for a visualization of the agreement levels for some specific high-importance aliases and tokens. The values in that matrix are the number of raters who assigned the given alias/classification to the token (n_{ij} in Eqs. 2,3). Recall that not all raters necessarily give a tag for every token.

In general, we see that agreement starts very high, when participants have tagged only a few tokens, and decreases as time goes on (although it remains high overall). This is likely due to the fact that the tool presents tokens in decreasing order of importance. Participants may be more likely to agree on how to classify the earlier, more important words, than they are on later words that occur more infrequently in the MWOs.

Another potential explanation for the dropoff in agreement is synonym cases. As can be seen in Figure 4, disagreement is most common in cases where different aliases have been assigned to words with the same meaning (e.g., "hose" and "line"). These pairs of words occur frequently in the data, especially in the important token set, which could help explain the differences in patterns between the full set of tokens and the reduced, popular set.

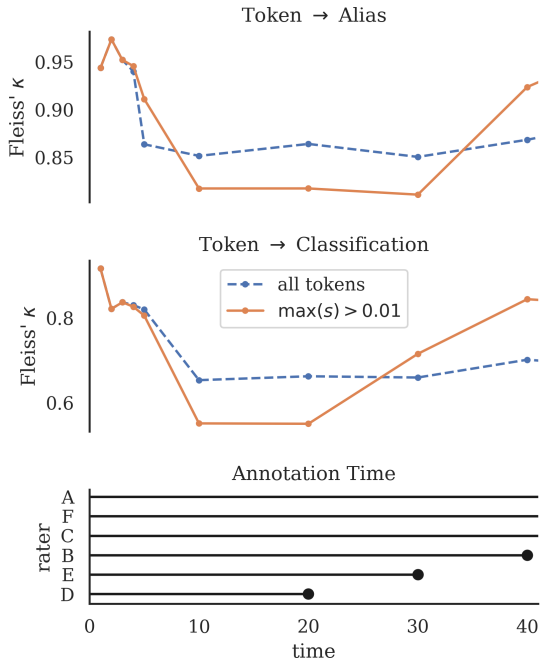


Figure 3. Agreement trends over time. Top: Alias agreement. Middle: Classification agreement. Bottom: Time progression of the task, showing participants still actively tagging.

Interestingly, we see in Figure 3 that agreement improves (especially for the important tags) after the participants who tagged for a shorter amount of time had finished. This trend points to how individual differences in annotators can have an impact on the tags assigned to tokens and the level of agreement reached.

3.3. Classification Agreement

In addition to measuring agreement on the aliases assigned to tokens, we also calculated κ_{Fleiss} for the classification of tokens into the concept categories used by Nestor. Inter-rater agreement at 30 min for token classification was 0.66, which indicates a good level of agreement, although lower than the other values of κ_{Fleiss} we measured.

As with the alias analysis, we additionally considered the participants' agreement on the classification of the most popular tokens. Here, κ_{Fleiss} was 0.72, which again indicates good agreement. See the right side of Figure 4 for a visualization of the classification agreement for a selection of the popular tokens.

We also repeated the analysis of changes in agreement over time. These results are shown in the middle section of Figure 3. Again, the figure shows these trends for both the full

Hastings, Emily; Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda.
"Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

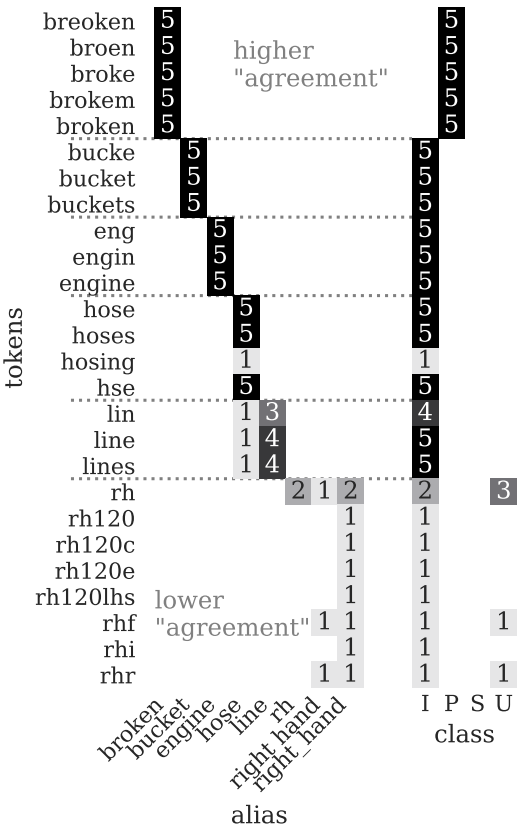


Figure 4. Matrix of rater agreement for a selection of high-importance tags and their corresponding tokens.

set of tokens and the reduced set of important tokens. The general trends are the same as described previously for the alias analysis.

One notable observation about the classification results is that the category which seemed to cause the most disagreement was "Unknown." We saw that many of the tokens had at least one participant label them as "Unknown," however these labels did not achieve the same kind of consensus that those given for the "Item" category did.

4. DISCUSSION

In this study, we examined the agreement behavior of multiple independent experts tagging MWO data for analysis. We found that the annotators were able to achieve high levels of agreement ($\kappa_{Fleiss} = 0.85$ for alias assignment and $\kappa_{Fleiss} = 0.66$ for concept classification on the set of all tags).

This finding has two main implications. First, since the aliases and labels assigned by a single expert are similar to those assigned by the entire group of experts, users of tools like Nestor can be reasonably confident that a single user can produce a set of valid tags that can be used in future analysis. Second, in the case that a single user is not tagging the entire dataset, users can again be reasonably confident that having multiple taggers will not compromise the quality of the set of tags produced.

However, our results also indicate a few opportunities to further improve agreement and the quality of the tags produced. We observed the following potential modes of disagreement:

1. Synonym token sets lead to split decisions among annotators, depending on domain ambiguity *See "hose" vs. "line" in Fig. 4. Some prefer retain specificity, while others tend to generalize.*
2. Token shorthand, ambiguity, and abbreviation is likely to be classified as "Unknown" by a subset of users. *This increases for tokens with many low-frequency variants, that not all users reach or understand (see "rh" in Fig. 4).*
3. Compound or multi-token aliases can hard to classify consistently. *For instance, some users retain "right hand" as two distinct concepts put together, while others chose "right_hand" as a single conceptual modifier.*

It is important to note the role that domain familiarity might be playing in the occurrence of these observed disagreements. As seen in Fig. 2, the speed of annotation, differs significantly among annotators. Additionally, the agreement level increased dramatically with the end of D and E annotators' participation (Fig. 3). This must be kept in mind, as such annotations performed in the field will likely come from analysts needing to prepare data, but not necessarily having

Hastings, Emily; Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda.
"Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

domain-specific expertise needed for high-agreement, reliable tagging.

In all of these modes, performing the tagging task in a collaborative environment that can offer real-time feedback, rather than tagging individually, could help resolve disagreements. For example, a user might be able to see how previous annotators tagged a token before finalizing their own decision, which may lead to greater convergence. This is especially true if the hypothesized familiarity differences are at play.

This concept of simultaneous groups of users arriving at a shared vocabulary at the most relevant level of abstraction for information retrieval is central to the idea of a *folksonomy* (Peters, 2009). These "folk"-taxonomies are built on the idea that annotation should reflect the vocabulary of a user-base (here, the technicians and operators storing information for future retrieval in MWOs). Folksonomies allow for convergence on a shared vocabulary in collaborative settings, and can facilitate communication between users, which is an ideal scenario for time-constrained practitioners on the shop-floor.

A limitation of this study is that we considered only one method of measuring agreement. By using κ_{Fleiss} we measured what could be called the "nominal agreement" between tokens and tags; i.e., did the exact label assigned to this token by Expert A match the exact label assigned by Expert B? An alternative way of conceptualizing agreement, particularly for aliases, would be to focus less on the exact alias assigned to a token, but rather on the group of tokens assigned the same alias, regardless of what that alias is.

One way to implement this more "topological" agreement measure would be to frame the annotation process as a graph and quantify disagreement using metrics like Graph Edit Distance (GED). Formulating crowd-sourced tagging efforts as graphs has a long and rich history in the folksonomy literature, where ranking the quality of folksonometric annotations through topology has been done with success previously (Hotho, Jäschke, Schmitz, & Stumme, 2006).

For instance, we could view an annotation session as the construction of two bi-partite graphs of the form $G = (V, E)$, where the vertex sets V can be split into two disjoint sets (e.g. the token nodes and the alias nodes, or the alias nodes and the classification nodes). The edges E in a bipartite graph only exist between sets; e.g. an "edge" between each of several tokens and their representative tag means that the tag is synonymous with all of its connected tokens.

In this way, such a bipartite graph would represent the annotations made by a user in a session of the experiment. The graph edit distances between each user's annotations would be high if the users disagreed on the the set of edges or the size of each node-set. This allows more flexibility in user annotation naming, valuing pattern similarity instead. Perhaps more interestingly, these GEDs could be quickly disaggregated and updated in realtime, with user similarity scores estimated over many time-steps for any of a number of local subgraphs, just as e.g. Eksombatchai et al., 2018 do for recommending pins/boards in the Pinterest "bipartite graph."

5. CONCLUSIONS AND FUTURE WORK

In this work we presented a preliminary study examining the agreement behavior of multiple isolated experts tagging MWO data. The results of the study have implications for implementing the tagging technique for MWO data analysis in authentic maintenance contexts: the annotators had high levels of agreement, suggesting that tagging by a single expert or by multiple experts are both feasible approaches. In addition, we identified potential opportunities for improvement of the tagging technique and tools that implement it. For example, performing the tagging task in a collaborative environment supporting real-time feedback could further improve the level of agreement achieved.

Domain knowledge could be an important factor in the level of agreement reached by multiple taggers. For example, users with less manufacturing experience may be more likely to tag concepts as "Unknown" than more experienced users, who might possess the background to apply a more appropriate classification, leading to a lower level of agreement between multiple taggers. Future work can further tease apart the effects of prior domain or tagging experience on agreement and investigate workflows that explicitly utilize differing levels of experience, such as those in which more experienced users train novices.

It is also possible that characteristics of the MWO dataset, such as complexity or domain, could impact the agreement among taggers. Future work could further investigate these factors' effect on agreement and whether our results generalize to other datasets.

Finally, the information that is provided from the MWO tagging needs to be incorporated into the maintenance decision workflows. This analysis is ongoing and will be further studied in future work.

ACKNOWLEDGMENT

Thanks to Dr. William Bernstein (NIST), Dr. Madhusudanan N (NIST), Mr. Michael Hoffman (NIST), Mr. Drew Georgiades (UWA), Ms. Emily Low (UWA), and Mr. Toby Griffiths (UWA) for their efforts annotating MWOs.

DISCLAIMER

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

Hastings, Emily; Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda.
"Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

REFERENCES

Ambati, V., Vogel, S., & Carbonell, J. G. (2010). Active learning and crowd-sourcing for machine translation. In *LREC* (Vol. 1, p. 2).

Brew, A., Greene, D., & Cunningham, P. (2010). Using crowdsourcing and active learning to track sentiment in online media. In *ECAI* (pp. 145–150).

Brundage, M. P., Kulvatunyou, B., Ademujimi, T., & Rakshith, B. (2017). Smart manufacturing through a framework for a knowledge-based diagnosis system. In *ASME 2017 12th international manufacturing science and engineering conference* (pp. V003T04A012–V003T04A012).

Brundage, M. P., Morris, K., Sexton, T., Moccozet, S., & Hoffman, M. (2018). Developing maintenance key performance indicators from maintenance work order data. In *ASME 2018 13th international manufacturing science and engineering conference* (pp. V003T02A027–V003T02A027).

Eksombatchai, C., Jindal, P., Liu, J. Z., Liu, Y., Sharma, R., Sugnet, C., ... Leskovec, J. (2018). Pixie: A system for recommending 3+ billion items to 200+ million users in real-time. In *Proceedings of the 2018 world wide web conference* (pp. 1775–1784).

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Folkrank: A ranking algorithm for folksonomies. In *Workshop information retrieval 2006 of the special interest group information retrieval.*

Hsueh, P.-Y., Melville, P., & Sindhwani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing* (pp. 27–35).

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93. Retrieved from `http://www.jstor.org/stable/2332226`

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge university press.

Mackworth, N. H., et al. (1950). Researches on the measurement of human performance. *Researches on the Measurement of Human Performance.*(268).

McCreadie, R. M., Macdonald, C., & Ounis, I. (2010). Crowdsourcing a news query classification dataset. In *Proceedings of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (cse 2010)* (pp. 31–38).

Nowak, S., & Rüger, S. (2010). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on multimedia information retrieval* (pp. 557–566).

Parasuraman, R. (1979). Memory load and event rate control sensitivity decrements in sustained attention. *Science*, *205*(4409), 924–927.

Parasuraman, R., & Davies, D. R. (1976). Decision theory analysis of response latencies in vigilance. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 578.

Peters, I. (2009). Folksonomies. indexing and retrieval in web 2.0. In (pp. 162–164). Walter de Gruyter.

Randolph, J. J. (2005, October). Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. In *Joensuu learning and instruction symposium.* Joensuu, Finland.

Sexton, T., Brundage, M. P., Morris, K., & Hoffman, M. (2017). Hybrid datafication of maintenance logs from AI-assisted human tags. In (p. 1-8). IEEE Big Data 2017.

Sexton, T., Hodkiewicz, M., Brundage, M. P., & Smoker, T. (2018). Benchmarking for keyword extraction methodologies in maintenance work orders. In *PHM society conference* (Vol. 10).

Smoker, T. M., French, T., Liu, W., & Hodkiewicz, M. R. (2017). Applying cognitive computing to maintainer-collected data. In *System reliability and safety (ICSRS), 2017 2nd international conference on* (pp. 543–551).

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254–263).

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Tarasov, A., Delany, S. J., & Cullen, C. (2010, October). Using crowdsourcing for labelling emotional speech assets. In *W3C workshop on emotion ML.* Paris, France. doi: 10.21427/D7RS4G

Hastings, Emily; Sexton, Thurston; Brundage, Michael; Hodkiewicz, Melinda.
"Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining."
Paper presented at Annual Conference of the Prognostics and Health Management Society 2019, Scottsdale, AZ, United States. September 21, 2019 - September 26, 2019.

Clustering and Representation of Time-Varying Industrial Wireless Channel Measurements

Mohamed Kashef*, Richard Candell†, and Yongkang Liu*
*Advanced Network Technologies Division †Intelligent Systems Division
National Institute of Standards and Technology (NIST)
Gaithersburg, Maryland, United States
Email: {mohamed.kashef, richard.candell, yongkang.liu}@nist.gov

***Abstract*—The wireless devices in cyber-physical systems (CPS) play a primary role in transporting the information flows within such systems. Deploying wireless systems in industry has many advantages due to lower cost, ease of scale, and flexibility due to the absence of cabling. However, industrial wireless deployments in various industrial environments require having the proper models for industrial wireless channels. In this work, we propose and assess an algorithm for characterizing measured channel impulse response (CIR) of time-varying wireless industrial channels. The proposed algorithm performs data processing, clustering, and averaging for measured CIRs. We have deployed a dynamic time warping (DTW) distance metric to measure the similarity among CIRs. Then, an affinity propagation (AP) machine learning clustering algorithm is deployed for CIR grouping. Finally, we obtain the average CIR of various data clusters as a representation for the cluster. The algorithm is then assessed over industrial wireless channel measurements in various types of industrial environments. The goal of this work is to have a better industrial wireless channels representation that results in a better recognition to the nature of industrial wireless communications and allows for building more effective wireless devices and systems.**

***Index Terms*—industrial wireless, wireless systems deployment, cyber-physical systems, wireless channel modeling, clustering, affinity propagation, channel impulse response**

I. Introduction

In future manufacturing systems, wireless communications technology plays an important role in achieving flexibility and scalability through allowing the communications between larger numbers of sensors and actuators and allowing more flexible mobility of equipment. The use of wireless communications in factory automation faces various challenges including the delay and reliability requirements, and the harsh industrial radio frequency (RF) environments [1]. This is due to the effects of noise and interference resulting from the broad operating temperatures, heavy machinery, and vibrations [2]. Moreover, the time-varying effects of moving objects may degrade the performance significantly if not considered [3]. As a result, knowledge of the wireless channel characteristics is essential for designing industrial wireless networks (IWNs).

Many questions are often asked about the best approach to be taken to ensure reliable operations of IWNs [4]. The initial step for industrial wireless deployment is defining objectives clearly by listing the enterprise goals prior to embarking on a wireless enhancement within a factory [5]. The deployment life cycle for an IWN is a generalized process that includes a business case with clearly stated objectives, a survey of the factory and technical requirements, candidate selection, solution design, and deployment [6]. Through IWN deployment, candidate solutions should be identified based on a merit-based selection process. A solution should then be developed, perhaps simulated, and tested. Hence, the radio frequency (RF) environments should be surveyed and modeled in order to assess various candidates and evaluate their performance.

The reliability of the wireless service is mainly affected by the multipath fading in industrial environments [7]. Such fading effects are caused by the multipath reflections on the metal surfaces and the moving objects within the environment. That results in correlated temporal variations in the transmitted signals over the industrial wireless channels [8]. These correlated variations can be captured through studying both the envelope variations stochastic model and the time-varying channel impulse response (CIR). Many works have argued that the fading distribution still follows Rician distribution even with the moving scatters in the environment [9]–[11]. However, this is only true with a large number of moving scatters which is not valid in field measurements [7]. On the other hand, obtaining an average CIR to model the correlated temporal variations cannot be performed over all the time-varying CIRs because of the different characteristics of these channels over time. However, these CIRs can be grouped, if possible, in order to obtain a CIR representation of each of these groups to model the correlated fading of the industrial wireless channels.

In order to evaluate various CIR averages of time-varying industrial environments, CIR matching can be used to obtain

Hany, Mohamed; Candell, Richard; Liu, Yongkang.
"Clustering and Representation of Time-Varying Industrial Wireless Channel Measurements."
Paper presented at 2019 IECON – the 45th Annual Conference of the IEEE Industrial Electronics Society, Lisbon, Portugal. October 14, 2019 - October 17, 2019.

groups of similar CIRs. The CIR is a time series that its shape is affected by the propagation delay and various multi-path components in the environment. Dynamic time warping (DTW) is a robust method used generally for calculating the similarity between various time series that are not aligned and affected by different compression and stretching effects [12]. On the other hand, Affinity propagation (AP) is a clustering algorithm which does not require the number of clusters in advance and offers an exemplar time series to represent each of the obtained clusters. Hence, it has been selected for CIR clustering and representation. Moreover, AP has been widely used in many applications [13] where it has proved to achieve the above mentioned goals adequately.

In this paper, we deploy the AP clustering scheme using DTW as a similarity measure for industrial wireless CIRs. The algorithm is assessed over the NIST measured CIRs in three different industrial environments [1]. Data preparation, before applying the clustering algorithm, has been performed including initial propagation delay alignment, filtering, and normalization. Later, similarity measures are evaluated, clusters are generated, and their average CIRs are presented.

The rest of the paper is organized as follows. Sec. II briefly discusses the related work. Sec. III describes the CIR data and identifies the required output of the clustering. Sec. IV explains in detail the proposed clustering scheme including the data preparation, similarity measure calculation, and cluster representation. Sec. V shows the obtained results. Finally, Sec. VI concludes the paper.

II. Related Work

Many measurement campaigns have been performed for various industrial environments to capture the characteristics of industrial wireless channels. Examples of these campaigns and their findings can be found in [1], [14]–[21] and the references therein. The National Institute of Standards and Technology (NIST) conducted RF propagation measurements at three selected sites of different classes of industrial environments. The CIRs for various measurement points are collected and used to obtain various metrics such as the path loss, delay spread, and K factor for various industrial wireless settings [1]. The focus of the works, [14]–[17], was studying the path loss and the envelope variation with various industrial environments. It was shown that heavy clutters lead to increased path loss and the envelope in line-of-sight (LOS) environments follows a Rician distribution. In [18]–[21], the power delay profiles of the measured channels are analyzed to optimize the transmitted signals and to study the CIR distribution. It was shown that the Saleh-Valenzuela model can describe CIRs in certain industrial factory halls as in Ref. [21]. However, the average CIR or the power delay profile to characterize industrial environments is only evaluated by a few works in the literature and for a stationary setting of the industrial environment.

DTW has been applied for various applications such as face detection, clustering, and many others [22]–[24]. Moreover, DTW is shown to improve the performance of time series clustering processes [25], [26].

By studying the similarity between CIRs, groups of homogeneous CIRs can be created and averaged in order to better represent their characteristics. Time series clustering is the process of finding the most homogeneous groups that are as distinct as possible from other clusters [27]. Various categories of clustering can be used in time series clustering including feature-based and shape-based clustering [28]. In feature-based clustering, a set of features are extracted from the time series and the clustering algorithm is performed over these features [29]. Various approaches for shape-based time series clustering can be used after defining the adequate similarity measure. Examples of clustering algorithms include the classical K-means algorithm and the K-Nearest Neighbor (KNN) algorithm [30]. Both these algorithms need the number of the clusters to be specified in advance which cannot be done in the CIR clustering problem.

III. Problem Formulation

In this section, we describe the available data and measurement sites. Three industrial sites with different characteristics are included in this study. We then focus on the goal of this work where we relate the measurement characteristics to the clustering process and the cluster representation. We also clarify the challenges in the data representation and the need for machine learning clustering in order to obtain the required data representation.

A. Measurement Environments

Measurements are taken at various industrial locations with different properties in order to study the effects of these environments on the RF propagation. A brief description is given in this work to motivate the proposed scheme and assess its performance. However, the details of the measurements can be found in [1].

The first environment is an automotive factory with dimensions of over 400 m x 400 m and the height is approximately 12 m. It contains both machines in open areas and enclosed spaces which are used to store factory inventory of small parts and tools. The factory was dense with tall metallic machines and concrete walls. The second environment is a steam plant with large machinery and overhead obstructions. The boiler section of the factory area was 20 m x 80 m with a height of 7.6 m. The walls are made of metal, concrete, and glass. Finally, measurements were taken in a small machine shop with outer dimensions of 25 m x 50 m and a height of 7.6 m.

B. Measured Data

The CIR data, which was collected in the various industrial environments, was generated by deploying a channel sounder. The channel sounder supports a 250 MHz instantaneous bandwidth and was used at a center frequency of 2.25 GHz. The center frequency was selected to have no interference from local wireless devices at the 2.4 GHz band while having similar propagation characteristics to this band. The sounder is

Hany, Mohamed; Candell, Richard; Liu, Yongkang.
"Clustering and Representation of Time-Varying Industrial Wireless Channel Measurements."
Paper presented at 2019 IECON – the 45th Annual Conference of the IEEE Industrial Electronics Society, Lisbon, Portugal. October 14, 2019 - October 17, 2019.

synchronized by having two rubidium clocks at the transmitter and receiver which ensure that drift between samples is small enough for accurate delay resolution of the multipath components and also allows for measurement of the absolute timing between transmitter and receiver. Mobile measurements were conducted by moving the RF receiver along a planned route at a constant speed. By taking different routes and reaching various locations in the industrial environments, it was possible to represent many classes of industrial channels.

The sounder produced a pseudo-noise (PN) sequence which was processed using cross-correlation calculations to produce the CIR for each record. These CIRs were stored with their associated Cartesian coordinates and time information. Then, the significant samples were kept by using a threshold criterion which required retaining samples 10 dB above the computed noise floor and within 30 dB from the CIR peak power. Finally, an iterative Grubbs test was used to detect single sample outliers and remove them using the code provided in [31].

C. Clustering Problem

The purpose of the proposed algorithm is to divide any set of CIRs into different groups, with highest within-group similarities and lowest among-groups similarities. Although this is the general purpose of any clustering problem, this problem includes unaligned time series clustering to an unknown number of clusters to obtain a representative CIR for each cluster. The input data is a set of thousands of CIRs at each industrial environment with each of these CIRs having thousands of samples with a 4 ns sampling rate. These CIRs are measured at different locations and have different propagation delays. The CIRs are also affected by various time-varying RF reflectors. Hence, the similarity measure should capture the shape of the CIR including time shifts, compression, and stretching. Depending on the environment characteristics, the number of output groups can be different in each location.

IV. AP-based CIR Clustering

In this section, we explain the various steps performed to achieve cluster representation for CIR measurements using DTW distance measure and AP clustering scheme. The process block diagram is shown in Fig. 1. We investigate the functions of each block in the following subsections.

A. Data Cleaning, Alignment, and Normalization

The initial stage is the data preparation where the measured CIRs are processed to be ready for the following stages. The details of the data preparation are shown in [1]. We start by removing statistical outliers of the CIR data through an iterative Grubbs test to detect outliers and remove them using the code provided in [31].

The initial alignment is performed to remove the propagation delay impact on the CIR data and to facilitate the process of calculating the DTW distance measure. The CIR start is determined by a threshold criterion in which the CIR starts by the first sample of 10 dB above the noise floor and 30 dB from the CIR peak power. Although initial alignment is performed, DTW distance measure is still needed to capture more precise distance among CIRs.

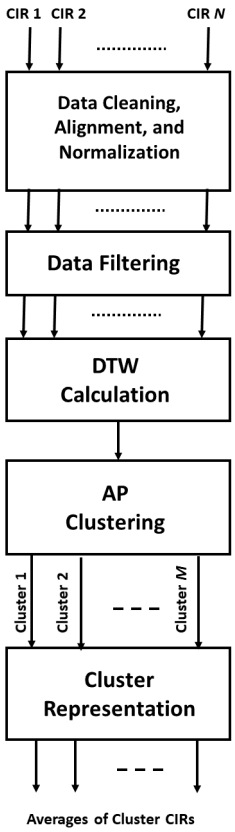


Fig. 1. Algorithm Block Diagram.

The final phase of this stage is the CIR energy normalization. In this phase, the amplitude of the CIR samples is scaled in order to have a total of unit energy. The CIR discrete time series is denoted by $x(t)$. The normalized version is evaluated as follows

$$\hat{x}(t) = \frac{x(t)}{\sqrt{\sum_{\tau=1}^{T} x^2(\tau)}}, 1 \leq t \leq T, \quad (1)$$

where t is the sample index and T is the total number of samples per CIR.

B. Data Filtering

A low-pass filter is used in order to remove the impact of high-frequency changes in the CIR which mainly results from erroneous samples in the CIRs due to hardware-related issues. An example of the filter impact on the CIRs is shown in Fig. 2. We have deployed a Butterworth low pass filter with cut off frequency 125 MHz.

Hany, Mohamed; Candell, Richard; Liu, Yongkang.
"Clustering and Representation of Time-Varying Industrial Wireless Channel Measurements."
Paper presented at 2019 IECON – the 45th Annual Conference of the IEEE Industrial Electronics Society, Lisbon, Portugal. October 14, 2019 - October 17, 2019.

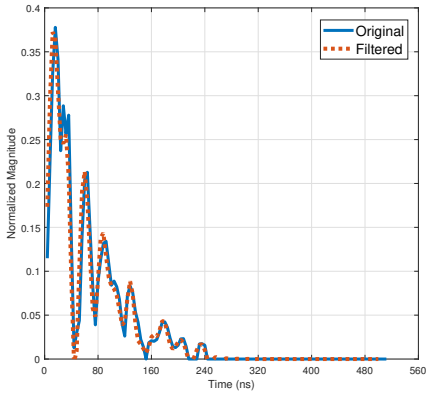


Fig. 2. Example showing the effect of data filtering on CIR.

C. Dynamic Time Warping Calculation

DTW is a non-linear measure to define a distance between two time series [26]. The two series $x_1(t)$ and $x_2(t)$ are defined by their samples $x_1(1), x_1(2), ..., x_1(T)$ and $x_2(1), x_2(2), ..., x_2(T)$, respectively. To evaluate the DTW distance between them, we start by building a T x T matrix where the (i, j)th element of the matrix corresponds to $d(i, j) = (x_1(i) - x_2(j))^2$. To find the distance between the two time series, the best match between the samples is evaluated by finding the path through the matrix to minimize the distance between them. The minimum cumulative path between the two series is selected as the DTW distance as follows:

$$\text{DTW}(x_1(t), x_2(t)) = \min_{w \in P} \sqrt{\sum_{k=1}^{K} d_{w_k}}, \tag{2}$$

where P is the set of all possible paths, w_k is position (i, j) at the kth segment of the path, and hence d_{w_k} is evaluated by $d(i, j)$ for the corresponding position, and K is the path length.

D. Affinity Propagation Clustering

AP clustering is based on the message passing concept and it does not require prior determination for the number of clusters. At the beginning, each time series is regarded as a cluster exemplar where the time series compete to continue as exemplars while others join the clusters. The responsibility message r_{mn} represents the message from the mth time series to the candidate cluster center n to describe the appropriateness of the nth time series as a cluster center to the mth time series. However, the availability a_{mn} is the message from the nth time series to describe the appropriateness of the mth time series to join the cluster [13]. Using the DTW measure between the mth and nth time series, the messages are computed as follows

$$r_{mn} = DTW(x_m(t), x_n(t)) - \max_{n' \neq n}(a_{mn'} - DTW(x_m(t), x'_n(t))). \tag{3}$$

$$a_{mn} = \begin{cases} \min\{0, r_{nn} + \sum_{m' \notin \{m,n\}} \max(0, r_{m'n})\}, m \neq n, \\ \sum_{m' \notin \{m,n\}} \max(0, r_{m'n}), m = n. \end{cases} \tag{4}$$

E. Cluster Representation

Two schemes can be used for cluster representation in this work. First, CIR averaging is the scheme where all the CIRs in a cluster are averaged together given that these CIRs are already aligned in the first stage of the algorithm. Second, we can use the cluster exemplar to represent the corresponding cluster. The cluster exemplar is not exactly the average of the CIRs, but the data point that has been selected by all the time series in the cluster as the cluster exemplar during the iterations of the AP clustering scheme. In the results section, we use the CIR averaging to represent the resulting clusters.

V. Results

In this section, we present examples of cluster representations using the proposed algorithm. We deploy the proposed algorithm in three data sets at three different environments as shown in [1]. We have deployed the (fastdtw) algorithm for an efficient implementation of the DTW distance calculation [32]. We have used the Scikit-learn implementation for the AP clustering algorithm with maximum iterations of 200 and a damping coefficient of 0.5 [33]. In this paper, we evaluate the CIR averages at various clusters to characterize the channel. We compare these averages to the simple approach of averaging all the measured CIRs at a specific scenario to justify the importance of the proposed algorithm. Further, we discuss future extensions for the proposed algorithm.

The first set of results is generated using the measured CIRs at the automotive factory which is characterized by reflective metals, machine canyons, and overhead robot gantry systems. By averaging all the measured CIRs in the environment as shown in Fig. 3, the environment can be characterized as a LOS environment with a relatively large delay spread, which is not correct because the environment has many locations of non-LOS (NLOS) channels.

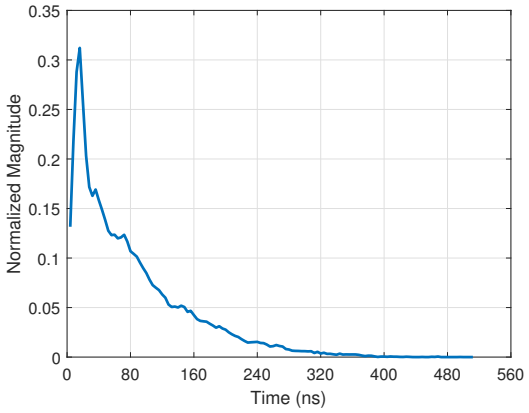


Fig. 3. The average CIR of all the measured CIRs at the Automotive factory.

Hany, Mohamed; Candell, Richard; Liu, Yongkang.
"Clustering and Representation of Time-Varying Industrial Wireless Channel Measurements."
Paper presented at 2019 IECON – the 45th Annual Conference of the IEEE Industrial Electronics Society, Lisbon, Portugal. October 14, 2019 - October 17, 2019.

In Fig. 4, we show the representation of the four resulting clusters by applying our proposed cluster representation algorithm. There are four groups of channels that can be described as follows: Cluster 1 represents NLOS CIRs with large delay spread, Cluster 2 is composed of LOS channels with other high amplitude reflections which still make a large delay spread, Cluster 3 has LOS CIRs with a small delay spread resulting from few and low magnitude reflections, and Cluster 4 has a larger delay spread than Cluster 3 but still is a LOS cluster with a high K-factor. Associating these CIR averages with the industrial environments can give a better representation for the CIR channels in these environments. We present the portions of the measured CIRs in the clusters in Fig. 5. This chart presents the percentage of CIRs that are represented by each of the cluster averages shown in Fig. 4. It can be shown that the NLOS CIRS are almost 41 % and hence the initial representation by averaging all the CIRs has mischracterized the environment by neglecting the impact of these 41 % of the CIRs.

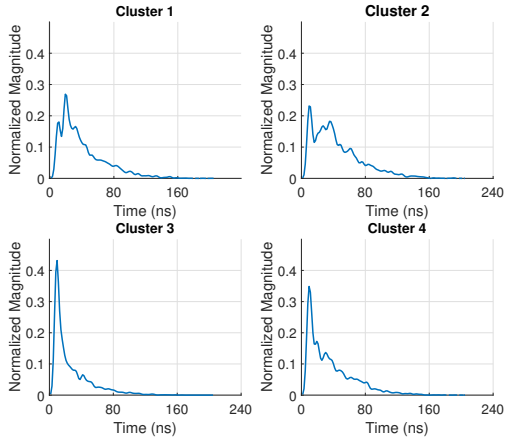


Fig. 4. The CIR cluster averages at the automotive factory.

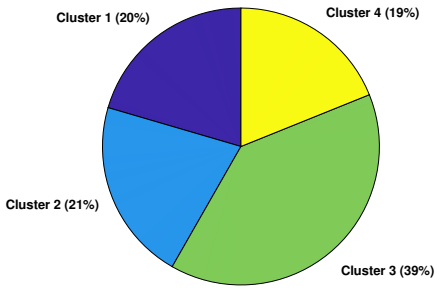


Fig. 5. The CIR clustering portions at the automotive factory.

Similarly in Figs. 6, 7, and 8, we show the CIR average in the steam plant for all the measured channels, the resulting cluster averages, and the clustering portions of the CIRs, respectively. The average CIR for all the measured channels does not offer much information about the environment. However, the resulting cluster average CIRs show that there are two clusters of NLOS behavior with high reflections, namely, Clusters 1 and 4. In the steam plant environment, there are more reflections with higher magnitude resulting from the existence of big metallic boilers on the factory floor. Clusters 2 and 3 contain LOS channels with high K-factor where the main difference is that the CIRs in Cluster 3 have a large magnitude reflection component very close to the LOS component which increases the delay spread slightly. Fig. 8 shows that Cluster 2 which does not have high reflections represents only 27 % of the total CIRs, while the rest of the CIR groups contain at least one high reflection component.

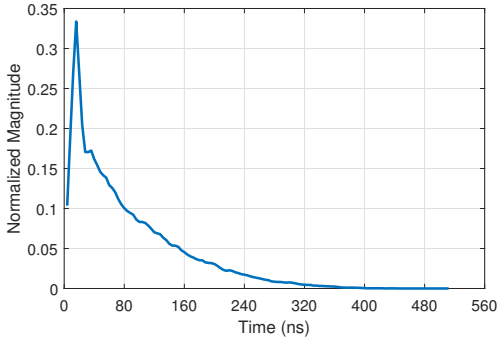


Fig. 6. The average CIR of all the measured CIRs at the steam plant.

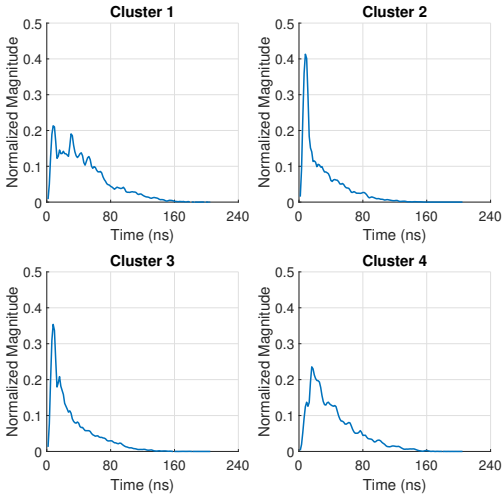


Fig. 7. The CIR cluster averages at the steam plant.

Hany, Mohamed; Candell, Richard; Liu, Yongkang.
"Clustering and Representation of Time-Varying Industrial Wireless Channel Measurements."
Paper presented at 2019 IECON – the 45th Annual Conference of the IEEE Industrial Electronics Society, Lisbon, Portugal. October 14, 2019 - October 17, 2019.

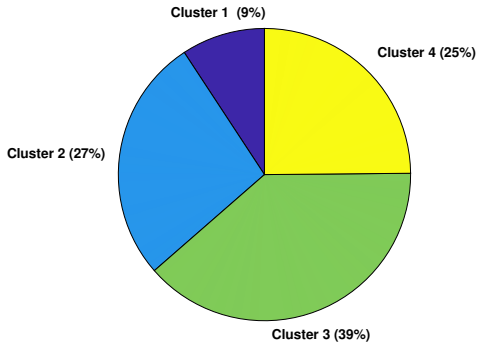


Fig. 8. The CIR clustering portions at the steam plant.

Finally in Figs. 9, 10, and 11, we have shown that the CIR average in the machine shop for all the measured channels, the resulting cluster averages, and the clustering portions of the CIRs, respectively. Although, the average CIR in Fig. 9 looks similar to the average CIR in both the automotive factory and the steam plant, the cluster representation shows that this machine shop has more chance to have LOS channels. A smaller delay spread compared to the other environments can be found in all the clusters including Cluster 4 with NLOS channels. From the clustering portions chart, the LOS CIRs represent 66 % of all the CIRs in the machine shop which is expected in such an environment with relatively small machines.

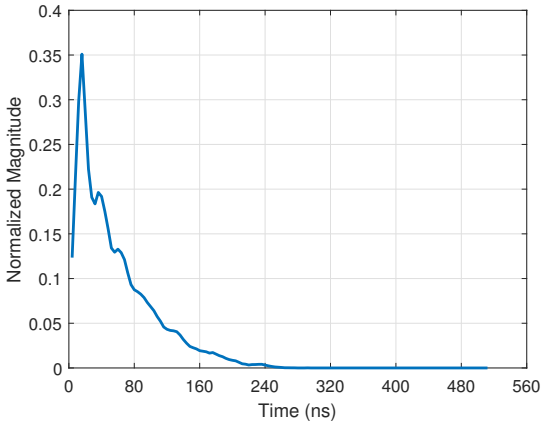


Fig. 9. The average CIR of all the measured CIRs at the machine shop.

VI. Conclusions

In this work, we have proposed an algorithm for time series clustering and representation that deploys a DTW distance measure among various industrial CIRs and an AP machine learning clustering scheme using the resulting DTW distances.

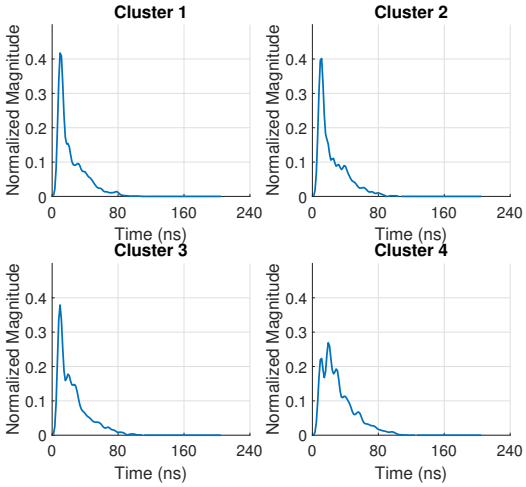


Fig. 10. The CIR clustering portions at the machine shop.

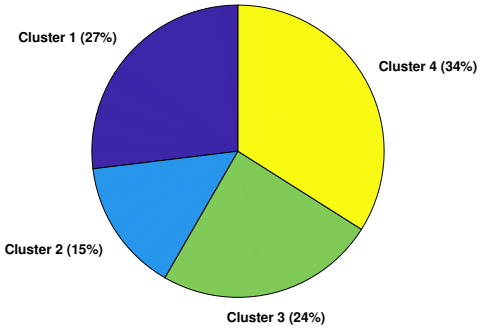


Fig. 11. The CIR cluster averages at the machine shop.

The algorithm helps in characterizing measured CIRs and obtaining suitable representation for reusing the measured CIRs. The proposed algorithm has been tested on measured data from three different industrial environments. We have shown that cluster representation can carry more information about the environments including the amount of reflectors and existence of LOS. We plan to work further in associating various cluster elements with their location information in order to have a better understanding of industrial environment impact on RF propagation. By having a better understanding on the behaviors of CIRs, more effective RF receivers can be designed and built for more reliable industrial wireless communications.

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental

Hany, Mohamed; Candell, Richard; Liu, Yongkang.
"Clustering and Representation of Time-Varying Industrial Wireless Channel Measurements."
Paper presented at 2019 IECON – the 45th Annual Conference of the IEEE Industrial Electronics Society, Lisbon, Portugal. October 14, 2019 - October 17, 2019.

procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

[1] R. Candell, C. Remley, J. Quimby, D. Novotny, A. Curtin, P. Papazian, G. Koepke, J. Diener, and M. Kashef, "Industrial wireless systems: Radio propagation measurements," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep., 2017. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.1951.pdf

[2] K. S. Low, N. Win, and M. J. Er, "Wireless sensor networks for industrial environments," in *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, vol. 2, Nov. 2005, pp. 271–276.

[3] M. Cheffena, "Propagation channel characteristics of industrial wireless sensor networks [wireless corner]," *IEEE Antennas and Propagation Magazine*, vol. 58, no. 1, pp. 66–73, Feb. 2016.

[4] R. Candell, M. Kashef, Y. Liu, K. B. Lee, and S. Foufou, "Industrial wireless systems guidelines: Practical considerations and deployment life cycle," *IEEE Industrial Electronics Magazine*, vol. 12, no. 4, pp. 6–17, Dec. 2018.

[5] Apprion Inc., "Building a business case for wireless at your industrial facility." [Online]. Available: https://www.automation.com/pdf_articles/apprion/Building_a_Business_Case_for_Wireless_WP_Oct_2011.pdf

[6] R. Candell, M. Kashef, K. B. Lee, Y. Liu, J. Quimby, and K. Remley, "Guide to industrial wireless systems deployments," Tech. Rep., National Institute of Standards and Technology (NIST), Apr. 2018. [Online]. Available: https://doi.org/10.6028/nist.ams.300-4

[7] Q. Zhang, Q. Zhang, W. Zhang, F. Shen, T. H. Loh, and F. Qin, "Understanding the temporal fading in wireless industrial networks: Measurements and analyses," in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct. 2018, pp. 1–6.

[8] B. Soret, M. C. Aguayo-torres, and J. T. Entrambasaguas, "Capacity with explicit delay guarantees for generic sources over correlated rayleigh channel," *IEEE Transactions on Wireless Communications*, vol. 9, no. 6, pp. 1901–1911, June 2010.

[9] X. Zhao, Q. Han, X. Liang, B. Li, J. Dou, and W. Hong, "Doppler spectra for f2f radio channels with moving scatterers," *IEEE Transactions on Antennas and Propagation*, vol. 64, no. 9, pp. 4107–4112, Sep. 2016.

[10] J. B. Andersen, J. O. Nielsen, G. F. Pedersen, G. Bauch, and G. Dietl, "Doppler spectrum from moving scatterers in a random environment," *IEEE Transactions on Wireless Communications*, vol. 8, no. 6, pp. 3270–3277, June 2009.

[11] T. Aulin, "A modified model for the fading signal at a mobile radio channel," *IEEE Transactions on Vehicular Technology*, vol. 28, no. 3, pp. 182–203, Aug. 1979.

[12] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1994, pp. 359–370. [Online]. Available: http://dl.acm.org/citation.cfm?id=3000850.3000887

[13] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007. [Online]. Available: https://doi.org/10.1126/science.1136800

[14] K. Zhang, L. Liu, C. Tao, K. Zhang, Z. Yuan, and J. Zhang, "Wireless Channel Measurement and Modeling in Industrial Environments," *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, no. 4, pp. 254–259, 2018.

[15] T. S. Rappaport and C. D. McGillem, "Uhf fading in factories," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 1, pp. 40–48, Jan. 1989.

[16] A. Miaoudakis, A. Lekkas, G. Kalivas, and S. Koubias, "Radio channel characterization in industrial environments and spread spectrum modem performance," in *2005 IEEE Conference on Emerging Technologies and Factory Automation*, vol. 1, Sep. 2005, pp. 87–93.

[17] C. Oestges, D. Vanhoenacker-Janvier, and B. Clerckx, "Channel characterization of indoor wireless personal area networks," *IEEE Transactions on Antennas and Propagation*, vol. 54, no. 11, pp. 3143–3150, Nov. 2006.

[18] E. Tanghe, W. Joseph, L. Martens, H. Capoen, K. V. Herwegen, and W. Vantomme, "Large-scale fading in industrial environments at wireless communication frequencies," in *2007 IEEE Antennas and Propagation Society International Symposium*, June 2007, pp. 3001–3004.

[19] B. Holfeld, D. Wieruch, L. Raschkowski, T. Wirth, C. Pallasch, W. Herfs, and C. Brecher, "Radio channel characterization at 5.85 ghz for wireless m2m communication of industrial robots," in *2016 IEEE Wireless Communications and Networking Conference*, April 2016, pp. 1–7.

[20] J. Ferrer-Coll, P. Angskog, J. Chilo, and P. Stenumgaard, "Characterisation of highly absorbent and highly reflective radio wave propagation environments in industrial applications," *IET Communications*, vol. 6, no. 15, pp. 2404–2412, Oct. 2012.

[21] J. Karedal, S. Wyne, P. Almers, F. Tufvesson, and A. F. Molisch, "Statistical analysis of the uwb channel in an industrial environment," in *IEEE 60th Vehicular Technology Conference, 2004. VTC2004-Fall. 2004*, vol. 1, Sep. 2004, pp. 81–85.

[22] H. Sahbi and N. Boujemaa, "Robust face recognition using dynamic space warping," in *Proceedings of the International ECCV 2002 Workshop Copenhagen on Biometric Authentication*. Berlin, Heidelberg: Springer-Verlag, 2002, pp. 121–132. [Online]. Available: http://dl.acm.org/citation.cfm?id=645307.649122

[23] W.-D. Chang and C.-H. Im, "Enhanced template matching using dynamic positional warping for identification of specific patterns in electroencephalogram," *Journal of Applied Mathematics*, vol. 2014, pp. 1–7, 2014. [Online]. Available: https://doi.org/10.1155/2014/528071

[24] H. C Huang and B. Jansen, "EEG waveform analysis by means of dynamic time-warping," *International journal of bio-medical computing*, vol. 17, pp. 135–44, Oct. 1985.

[25] J. Hu, B. Ray, and L. Han, "An interweaved HMM/DTW approach to robust time series clustering," in *18th International Conference on Pattern Recognition (ICPR06)*. IEEE, 2006. [Online]. Available: https://doi.org/10.1109/icpr.2006.257

[26] G. Chen, Q. Wei, and H. Zhang, "Discovering similar time-series patterns with fuzzy clustering and DTW methods," in *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*. IEEE. [Online]. Available: https://doi.org/10.1109/nafips.2001.944404

[27] P. Esling and C. Agon, "Time-series data mining," *ACM Comput. Surv.*, vol. 45, no. 1, pp. 12:1–12:34, Dec. 2012. [Online]. Available: http://doi.acm.org/10.1145/2379776.2379788

[28] T. W. Liao, "Clustering of time series data—a survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, Nov. 2005. [Online]. Available: https://doi.org/10.1016/j.patcog.2005.01.025

[29] M. G. Baydogan, G. Runger, and E. Tuv, "A bag-of-features framework to classify time series," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2796–2802, Nov. 2013. [Online]. Available: https://doi.org/10.1109/tpami.2013.72

[30] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*, 1st ed. Chapman & Hall/CRC, 2013.

[31] B. Shoelson, "Matlab code deleteoutliers," Sep. 2003. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/3961-deleteoutliers

[32] S. Salvador and P. Chan, "FastDTW: Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, Oct 2007. [Online]. Available: http://doi.org/10.3233/IDA-2007-11508

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

Hany, Mohamed; Candell, Richard; Liu, Yongkang.
"Clustering and Representation of Time-Varying Industrial Wireless Channel Measurements."
Paper presented at 2019 IECON – the 45th Annual Conference of the IEEE Industrial Electronics Society, Lisbon, Portugal. October 14, 2019 - October 17, 2019.

Compositional models for power systems

John S. Nolan[1], Blake S. Pollard[2], Spencer Breiner[2], Dhananjay Anand[2], and Eswaran Subrahmanian[3]

[1]University of Maryland, College Park, MD, USA

[2]National Institute of Standards and Technology, Gaithersburg, MD, USA

[3]Carnegie Mellon University, Pittsburgh, PA, USA

The problem of integrating multiple overlapping models and data is pervasive in engineering, though often implicit. We consider this issue of model management in the context of the electrical power grid as it transitions towards a modern 'Smart Grid.' We present a methodology for specifying, managing, and reasoning within multiple models of distributed energy resources (DERs), entities which produce, consume, or store power, using categorical databases and symmetric monoidal categories. Considering the problem of distributing power on the grid in the presence of DERs, we show how to connect a generic problem specification with implementation-specific numerical solvers using the paradigm of categorical databases.

1 Introduction

The modeling of complex systems, engineered or natural, entails certain generic challenges: the existence and interaction of multiple models, multiple algorithms, and multiple implementations. This paper presents a methodology rooted in category theory to manage this complexity, concretized via a model-driven engineering approach to designing a modern electrical grid, dubbed the 'Smart Grid.'

The existing grid architecture is characterized by dedicated large-scale, centralized generation and distributed, downstream consumption. Moving towards an architecture with increased distributed generation will have a profound impact on how the grid is managed: end users will no longer be dedicated

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

consumers, but will shift between consuming and producing power. One key to enabling this transition is the management and modeling of distributed energy resources (DERs), generic devices that can consume, produce, or store power.

The notion of DER is meant to provide an abstraction or characterization summarizing the essential properties of a wide array of different energy resources, e.g. photovoltaic systems, batteries, conventional loads, and so on. The issue is that no uniform abstraction exists. Different stakeholders utilize different abstractions for different purposes. In addition, these meta-models must evolve as new technologies emerge.

Coupled with control mechanisms, DERs provide a number of ancillary services to consumers and grid operators: voltage control, reducing peak loads, demand response, etc. [15, 18]. Aggregations of heterogeneous DERs provide the abstraction through which such collections participate in the overall power system and energy markets [4].

There is a large body of work concerned with the use of model transformations in the context of model management and model-driven engineering, e.g. [13, 24]. A number of approaches utilize category theory, recognizing the natural mathematical framework it provides for reasoning about models, their semantics, and structure-preserving transformations among models [5, 25].

We tackle the problem of specifying, relating, and transforming models using the functorial data model advocated in [3, 8, 19, 20, 22] as well as its computational implementation in the CQL tool. In the functorial data model, database schemas are interpreted as finite presentations of categories. Instances of a database schema correspond to Set-valued functors out of the associated category. Some subtleties arise when working with computational data such as strings and integers, though we will not concern ourselves with these difficulties; see [20] for a thorough treatment.

Structure of this paper: In Section 2, we consider the problem of model specification and transformation for family of DER schemas and the functors among them. In Section 3, we show that one variant of a model in this family yields the objects of a symmetric monoidal category with DER aggregation as monoidal product. In Section 4, we consider the problem of distributing power in a grid where certain nodes correspond to aggregate collections of DERs, describing a procedure for translating among numerical solvers using database schemas, functors, and queries.

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

2 Categorical databases for model management

The family of models we present in this section share a common ancestor, the directed multigraph, henceforth graph, consisting of two entities, States and Transitions, together with two arrows, Source and Target, assigning source and target states to transitions:

Including identities and composites, this schema forms a category DiGraph. Functors from this category to Set form a category of instances.

2.1 A basic DER model

In our base model, DERs are viewed as graphs with operational states as nodes and transitions among those. Each state is assigned a feasible operating region (i.e. power demands / generation). In AC circuits, power is a complex-valued quantity $P + iQ$, where the real part P is referred to as real or active power and the imaginary part Q as reactive power. For now we restrict our attention to the case where operating regions are single points.

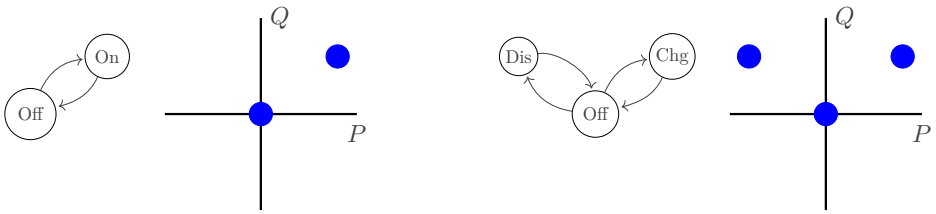
Our base DER model is described by the schema $\mathsf{DER}_{\mathsf{Base}}$

consisting of a graph together with two attributes for each state, P, Q: State $\to$ Float. Two instances of $\mathsf{DER}_{\mathsf{Base}}$ are depicted in Figure 1, showing a typical load (an HVAC, i.e. heating/cooling system) and a battery.

2.2 Model Translation via Functors

Depending on the analysis to be performed, this basic DER meta-model or schema extends to include additional information such as state of charge, virtual cost of transitions, location, etc. The functorial data model offers a robust collection of ways to translate between such models. Some of these models and the functors relating them are summarized in the following and in Figure 2.

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.



(a) HVAC (b) Battery

Figure 1: Two types of DERs and their associated demand profiles. With the chosen convention, states with positive real power P consume power, while states with negative P generate power.

In [1, 16], virtual costs are assigned to transitions representing the willingness/ability of a controllable DER to perform a certain transition. This leads to a new schema $\mathsf{DER}_{\mathsf{Cost}}$ with an additional attribute for transitions and obvious inclusion functor to it from the base DER model.

Attaching non-negative rates to transitions on a graph gives a Markov process, summarized by the schema Markov [14]. For now, we implement these rates as 'Floats,' tabling the discussion of constraints in CQL until Section 4.2. Quantifying variability of generation from renewable sources is a key issue when modeling DERs. One approach models DERs as Markov processes, giving a stochastic DER model $\mathsf{DER}_{\mathsf{Mark}}$.

Over long time scales, the steady state probabilities of such a model can be used to estimate energy production and other performance indices. In [12], this approach is utilized to evaluate reliability of small wind farm generation by assigning probabilistic transitions between operative and failed states and coupling this with a stochastic model of wind variability. This methodology is also applied to small hydro electric stations in [2]. Stochastic models of solar irradiance are also used to generate synthetic data for system design [26]. We summarize these stochastic models of weather in the schema Weather.

Functors $F\colon M \to N$ between database schemas give rise to adjoint triples of functors $\Sigma_F, \Delta_F, \Pi_F$ between the associated categories of instances, where $\Delta_F\colon N\text{–}\mathsf{Inst} \to M\text{–}\mathsf{Inst}$ and $\Sigma_F, \Pi_F\colon M\text{–}\mathsf{Inst} \to N\text{–}\mathsf{Inst}$. These functors are related to "uber-flower" queries $Q\colon M \to N$. Such a Q can be evaluated (as in other data models) to give a functor $\mathsf{eval}(Q)\colon M\text{–}\mathsf{Inst} \to N\text{–}\mathsf{Inst}$ or dually "coevaluated" to give a functor $\mathsf{coeval}(Q)\colon N\text{–}\mathsf{Inst} \to M\text{–}\mathsf{Inst}$. Using data migration functors and queries, along with CQL's ability to compute colimits of instances, offers

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

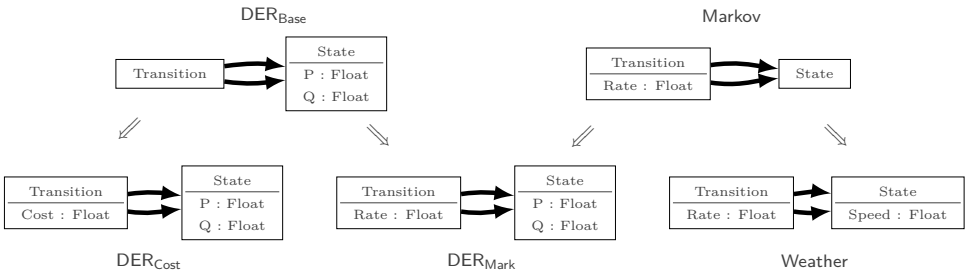


Figure 2: Multiple model schemas (boxes are objects, dark arrows are morphisms), connected by functors (the hollow arrows). On the left, DERs with costly transitions share a common underlying base model with a stochastic DER model. On the right, stochastic DER models and stochastic models of natural processes are both modeled using Markov processes.

a useful way to translate between the data associated to different models.

The models presented here are connected by inclusion functors, summarized in Figure 2. Even in the simple setting described here, the use of mappings connecting DER models enables reuse and validation of said models, while providing an extensible framework for model documentation. Implementing mappings between models which utilize the same method, such as Markov processes, enables the reuse of tools or methods, e.g. steady state solvers. In Section 4.4, we give a more detailed exposition on how functors between schemas along with their associated adjoint triples and queries can be used to connect models to tools within the paradigm of categorical databases.

In the next section we give a categorical treatment of DER aggregation, the process of taking collections of DERs and combining them into a single DER. For this we move to the setting of symmetric monoidal categories and consider demand regions as subsets of the complex plane.

3 Aggregation as a Symmetric Monoidal Product

Aggregation of DERs is the key to unlocking their potential to provide ancillary services such as peak shaving and voltage control. It is often third-party aggregators who act as intermediaries between utilities and customers, pooling resources and providing data integration and control strategies via the development of distributed energy resource management systems (DERMS) [7].

In this section, we focus on the basic problem of aggregating demand re-

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

gions and state spaces of DERs, presenting a symmetric monoidal category DER whose objects are DERs, and where aggregation serves as the tensor product. This model adds a reflexive property [23], i.e. mandatory self-edges for each node, to the underlying graphs of the DER models outlined previously. Morphisms in DER correspond to adjusting the level of granularity of the state space.

Definition 1. A ***distributed energy resource (DER)*** $\mathcal{D} = (S, T, s, t, r, d)$ consists of a graph $s, t \colon T \to S$, together with a function $r \colon S \to T$, satisfying $s \circ r = t \circ r = \mathrm{id}_S$, picking out an ***identity transition*** from each state to itself, and a function $d \colon S \to 2^{\mathbb{C}}$ assigning to each state $\sigma \in S$ a ***power demand region*** $d(\sigma) \subseteq \mathbb{C}$. For each state $\sigma \in S$ we write $1_\sigma = r(\sigma)$ and call 1_σ the identity transition of σ.

This definition is summarized by the diagram $T \underset{r}{\overset{s \quad t}{\rightleftarrows}} S \xrightarrow{d} 2^{\mathbb{C}}$.

Definition 2. A ***morphism of DERs*** $\phi \colon \mathcal{D} \to \mathcal{D}'$ consists of a pair of functions (ϕ_S, ϕ_T), where $\phi_S \colon S \to S'$ and $\phi_T \colon T \to T'$, such that for all $\tau \in T$, $\phi_S(s(\tau)) = s'(\phi_T(\tau))$ and $\phi_S(t(\tau)) = t'(\phi_T(\tau))$, and for all $\sigma \in S$, $\phi(1_\sigma) = 1_{\phi(\sigma)}$ and $d(\sigma) \subseteq d'(\phi_S(\sigma))$. Together with these morphisms (and the obvious identity morphisms and composition law), DERs form a category which we denote DER.

In short, a morphism of DERs is a homomorphism of the underlying graphs that acts as an inclusion of subsets on the demand regions for each state. Such morphisms can be used to translate between models of a DER, e.g. by adding more states or by merging states which are indistinguishable in the codomain model. An example of this is provided in Subsection 3.1.

Demands can be aggregated using Minkowski sums; see [6] for more details as well as [10] for an application to modeling the flexibility of DERs.

Definition 3. Given two subsets $X, Y \subseteq \mathbb{C}$, the ***Minkowski sum*** of X and Y is the set

$$X + Y = \{x + y : (x, y) \in X \times Y\} \subseteq \mathbb{C}.$$

Under this operation, $2^{\mathbb{C}}$ is a commutative monoid with unit $\{0\}$.

Definition 4. The ***aggregate*** of two DERs $\mathcal{D}$ and $\mathcal{D}'$ is the DER $\mathcal{D} \otimes \mathcal{D}' = (S \times S', T \times T', s \times s', t \times t', r \times r', d + d')$, where $d + d' \colon S \times S' \to 2^{\mathbb{C}}$ is defined by $(d + d')(\sigma, \sigma') = d(\sigma) + d'(\sigma') \subseteq \mathbb{C}$ for any $(\sigma, \sigma') \in S \times S$.

In short, the aggregate of two DERs is the categorical product of the underlying graphs (see [23] or [17, Proposition 3.3.9]), where each product state

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

is equipped with demand equal to the Minkowski sum of its factors. Observe that the reflexive property of individual DERs within an aggregate DER enables independent transitions.

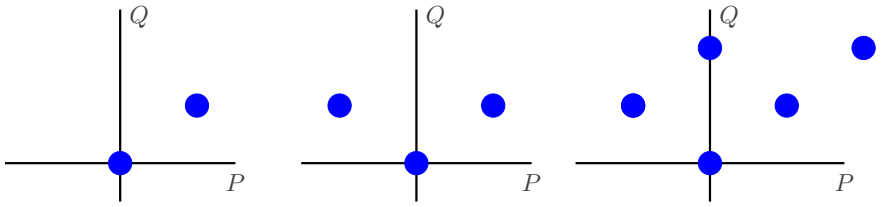
Aggregation extends easily to morphisms by taking Cartesian products of functions, so in this way we see $\otimes\colon \mathsf{DER} \times \mathsf{DER} \to \mathsf{DER}$ is a bifunctor. In fact, letting $\mathcal{I}$ denote the DER with one state σ, a single transition 1_σ, and power demand $d(\sigma) = \{0\} \subseteq \mathbb{C}$, it is not hard to show that DER is a symmetric monoidal category with tensor product $\otimes$ and unit $\mathcal{I}$. As a result, string diagrams can be used to reason about DERs and aggregation [9, 21].

3.1 Net Demand Quotient

When aggregating DERs, the state space grows rapidly. For operations at the distribution level, all that is relevant is the *net* power demand. Thus it is natural to mod out by an equivalence relation whereby states with identical power demand are identified. The following definition formalizes this notion.

Definition 5. Let $\mathcal{D}$ be a DER. Consider the equivalence relation $\sim$ on the states S of $\mathcal{D}$ where $\sigma \sim \sigma'$ if and only if $d(\sigma) = d(\sigma')$. This induces an equivalence relation $\approx$ on the edges T of $\mathcal{D}$ where $\tau \approx \tau'$ if and only if $s(\tau) \sim s(\tau')$ and $t(\tau) \sim t(\tau')$. We can define the ***net demand DER*** $\overline{\mathcal{D}}$ of $\mathcal{D}$ by $\overline{\mathcal{D}} = (S/\!\sim, T/\!\approx, \overline{s}, \overline{t}, \overline{r}, \overline{d})$, where $\overline{s}, \overline{t}, \overline{r}$, and $\overline{d}$ are defined in the obvious way.

The equivalence relation above gives rise to a DER morphism $\overline{(\ )}\colon \mathcal{D} \to \overline{\mathcal{D}}$ which identifies states with equal power demand and transitions among them. Composing this morphism with aggregation applied to a pair of DERs $\mathcal{D}$ and $\mathcal{D}'$ gives a DER $\overline{\mathcal{D} \otimes \mathcal{D}'}$ which only distinguishes states which differ in their net power demand.



(a) HVAC (b) Battery (c) HVAC-Battery

Figure 3: The demand profile hybrid or aggregate DER consisting of an HVAC system and a battery.

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

Any path in $\overline{\mathcal{D}}$ will give a set of paths in $\mathcal{D}$ traveling among DER states. We can then consider methods for selecting the 'best' or 'least-costly' sequence of DER transitions which accomplish some desired transition in net demand. This allows for dynamic tasking of DERs to accommodate demand fluctuations without requiring distribution level operators to have full knowledge of the details of a collection of DERs.

DER aggregation is typically done locally/regionally, interfacing with grid operators at the distribution or transmission level where the problem becomes matching generation with consumption while maintaining stable operating conditions. We now turn to the basic problem of distributing electricity through the grid so as to match production and consumption. specification and numerical solution of basic power flow problems using categorical databases.

4 Power Flow Problems

In this section, we show how to connect models with tools or solvers by describing the specification and numerical solution of basic power flow problems using categorical databases. This amounts to finding solutions to a set of nonlinear equations, the power flow equations, defined over a network or power flow graph:

Definition 6. A ***power flow graph*** consists of graph $s, t\colon E \to N$, together with functions $g, b\colon E \to \mathbb{R}$, assigning a ***conductance*** and ***susceptance*** to each edge. Nodes in the graph $n \in N$ are typically called ***buses***, while edges $e \in E$ are referred to as ***branches***. Conductance and susceptance are the real and imaginary parts of the complex admittance, a measure of the susceptibility of a branch to admitting current flow.

The variables of interest are the real and imaginary parts of the complex power $P + iQ$ and the magnitude and phase of the complex voltage $Ve^{i\theta}$, which we regard as partial functions $P, Q, V, \theta\colon N \to \mathbb{R}$. Buses are typed as PQ, PV, or $V\theta$ buses according to which pair of variables is regarded as fixed, see Figure 4. The remaining free variables are determined by solving the power balance equations.

Definition 7. The ***power balance equations*** [11] for a power flow graph are the $2|N|$ equa-

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

tions

$$P_i = V_i \sum_j V_j \left(\, g_{ij} \cos(\theta_i - \theta_j) + b_{ij} \sin(\theta_i - \theta_j) \, \right)$$
$$Q_i = V_i \sum_j V_j \left(\, g_{ij} \sin(\theta_i - \theta_j) - b_{ij} \cos(\theta_i - \theta_j) \, \right),$$

where we write $P_i := P(N_i)$ and $g_{ij} := g(E_{ij})$ etc. and each sum is taken over all buses adjacent to i.

We summarize the data needed to specify a power flow problem in a CQL schema in Figure 4, omitting attributes for simplicity.

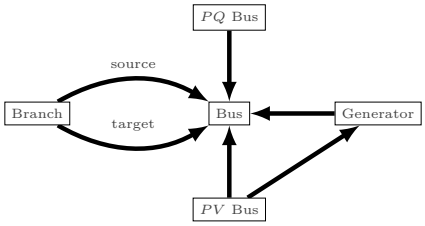


Figure 4: A schema describing a generic power flow problem. A PQ bus represents a typical load, whose real and reactive power demands are known and fixed, at any moment of time. All PV buses are viewed as having generators attached, producing constant power at a specific voltage. Slack buses are omitted for visual clarity.

Due in part to their non-linearity, solving the power flow equations is typically done numerically either using freely available software, commercial tools, or customized code. Such tools usually require specific solver parameters and use their own internal data structures.

4.1 Connecting to a Tool

MATPOWER is a commonly used power systems toolbox, implemented in MATLAB. The MATPOWER data format specifications are organized into tables in Appendix B of the MATPOWER manual [27]. We translate these specifications into MATPOWER-specific schemas in CQL. Figure 5 shows the resulting schemas representing both a power flow problem as well as an associated solver, e.g We characterize an iterative Newton-Raphson solver in terms of its required parameters such as tolerance, maximum number of iterations, etc.

Encoding the input problem specification, the output solution structure, as well as the solver parameters in database schemas enables systematic experimentation, i.e. varying inputs or parameters, while providing flexible and

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

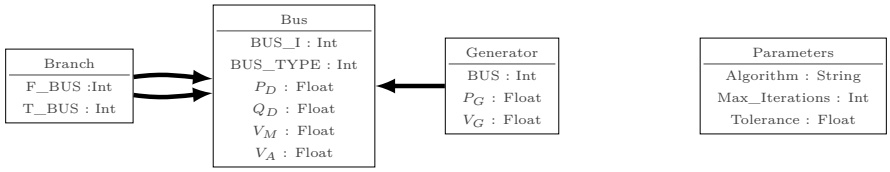
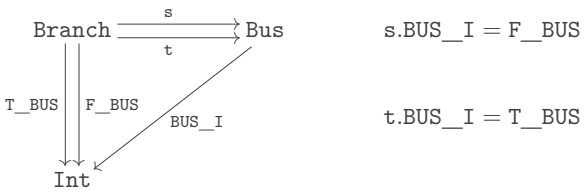


Figure 5: A MATPOWER power flow schema on the left, with solver parameters on the right. For simplicity we only show a few attributes for each entity. Attribute names are based on those in MATPOWER. Compare Figure 4, which was developed based on a reorganization of the schema on the left.

traceable documentation, i.e. storing just solutions or including the solver settings used in each run. The input and output features common to all solvers of a given type can be organized into a generic schema for solvers of that given type.

4.2 Constraints in CQL

CQL allows for the enforcement of constraints in the form of path equations. For example, consider the chunk of our MATPOWER schema:



The equations on the right enforce the constraint that the indexing of buses via BUS_I is consistent with the indexing of `T_BUS` and `F_BUS` of branches.

4.3 Connecting to DERs

To interface with a standard power flow problem, we place DERs at the relevant nodes of a power flow graph, treating each such node as a PQ bus. For each such bus we determine average P, Q values from the DERs at that node, for example by modeling the relevant DERs as Markov chains, as described in Subsection 2.1, and returning the sum of the expected steady-state P, Q values for each DER. This process is depicted in Figure 6 and implemented by the authors in a MATPOWER example.

This hybrid setup enables the exploration on the dependence of overall solutions to the power flow equations on the types and behaviors of DERs, e.g.

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

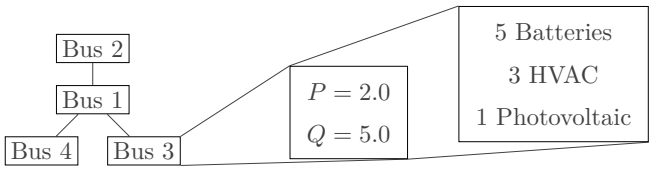


Figure 6: Incorporating data from a collection of DERs into a node in a power flow graph. Units and values of P, Q are arbitrary.

how stochasticity of distributed generation enters into overall power distribution. We now turn to the problem of connecting multiple tools or solvers.

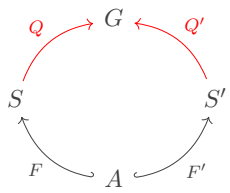
4.4 Connecting Tools

Modeling something as complex as the electrical grid typically involves collaborations among teams who may utilize a variety of tools or implementations, even for the same or similar problems. This creates a need for translation and validation among different solvers. We describe a procedure for accomplishing this task using techniques from the functorial data model, as presented in Subsection 2.1.

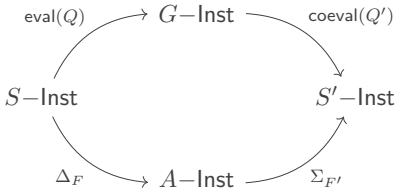
Figure 7 provides diagrams describing how to translate between solvers. Consider two solvers for some problem, represented by schemas S and S', e.g. the schema in Figure 5 and a schema for a solver with a different set of parameters. One can construct a generic solver schema G for the problem, e.g. that in Figure 4, along with queries $Q\colon S \to G$ and $Q'\colon S' \to G$, specifying which information is shared among the generic and specific instances. In this case, one should also define an auxiliary schema A for data which appears in both S and S' but not in G, as well as functors $F\colon A \to S$ and $F'\colon A \to S'$ inserting the data of A into both specific solver schemas.

These constructions give rise to functors between the associated categories of instances, as depicted in Figure 7b. For every instance I of S, we can obtain two instances of S' by applying these functors. These can be combined using a suitable colimit in S'–Inst to get a single instance of S' containing all possible data from S. Such a construction enables one to translate between the inputs, outputs, and parameters for the solver represented by S and the corresponding values for the solver represented by S'.

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.



(a) Transformation between different solver schemas for the same problem.



(b) Transformation between instances of different solver schemas.

Figure 7: Diagram depicting the transformation of instances for solver schema S to instances for solver schema S'. Black arrows are functors; red arrows are queries.

5 Conclusions and Future Work

This paper provides a window into our efforts to concretize the potential utility of a category theoretic viewpoint for problems dealing with multiple related models and tools in the context of power systems engineering. We saw that techniques and tools from categorical databases can readily be applied to specify and translate among various models, connecting those models to particular analysis tools, as well as connecting various tools themselves.

Further work is required to extend this category-theoretic modeling paradigm to other engineering domains as well as within power systems. What is desired is not a modeling framework which captures the full complexity of the today's grid, but rather a framework which enables the expedient exploration and evaluation of various possible future architectures and pathways to those. The need for such a modeling ecosystem is not unique to power systems.

Of particular relevance for future work in Smart Grid technologies are aspects of control and communication enabled by new devices such as Smart Meters and increased deployment of phasor measurement units (PMUs), devices which measure current, voltage, or phase across the grid. Managing this coupling of an information network with a physical power network presents ample opportunities for applied category theorists.

Lastly, further development of tools for specifying and modeling systems using category theory, e.g. CQL, is essential in terms of engagement with domains. Being able to point practitioners to a system they can get their hands on and play with goes a long way towards arriving at a useful common understanding.

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

Acknowledgements Source code for the examples discussed can be found at github:AQL_Powersystems. The authors would like to thank David Spivak and Ryan Wisnesky of Categorical Informatics Inc. for helpful discussions. This material is based upon work supported by the National Science Foundation under Grant No.1746077. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. JN was supported by the NIST SURF and NIST PREP programs. BP was supported by an NRC Postdoctoral Research Associateship.

References

[1] Andrey Bernstein, Lorenzo Reyes-Chamorro, Jean-Yves Le Boudec, and Mario Paolone. A composable method for real-time control of active distribution networks with explicit power setpoints. part I: Framework. *Electric Power Systems Research*, 125:254 – 264, 2015. ISSN 0378-7796. DOI: https://doi.org/10.1016/j.epsr.2015.03.023. URL http://www.sciencedirect.com/science/article/pii/S0378779615000905.

[2] C. L. T. Borges and R. J. Pinto. Small hydro power plants energy availability modeling for generation reliability evaluation. *IEEE Transactions on Power Systems*, 23(3):1125–1135, Aug 2008. ISSN 0885-8950. DOI: 10.1109/TPWRS.2008.926713.

[3] Spencer Breiner, Blake Pollard, and Eswaran Subrahmanian. Functorial model management. In *22nd International Conference on Engineering Design*, page To appear. The Design Society, 2019.

[4] Georgios Chalkiadakis, Valentin Robu, Ramachandra Kota, Alex Rogers, and Nicholas R Jennings. Cooperatives of distributed energy resources for efficient virtual power plants. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 787–794. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

[5] Zinovy Diskin and Tom Maibaum. Category theory and model-driven engineering: From formal semantics to design patterns and beyond. *Model-Driven Engineering of Information Systems: Principles, Techniques, and Practice*, page 173, 2014.

[6] Rida T Farouki, Hwan Pyo Moon, and Bahram Ravani. Minkowski geometric algebra of complex sets. *Geometriae Dedicata*, 85(1-3):283–315, 2001.

[7] Eric O'Shaughnessy Brittany Smith Jeffrey J. Cook, Kristen Ardani and Robert Margolis. Expanding pv value: Lessons learned from utility-led distributed energy resource aggregation in the united states. Technical report, Golden, CO, 2019.

[8] Michael Johnson, Robert Rosebrugh, and RJ Wood. Entity-relationship-attribute designs and sketches. *Theory and Applications of Categories*, 10(3): 94–112, 2002.

[9] André Joyal and Ross Street. The geometry of tensor calculus, I. *Advances in Mathematics*, 88(1):55 – 112, 1991. ISSN 0001-8708. DOI: https://doi.org/10.1016/0001-8708(91)90003-P. URL http://www.sciencedirect.com/science/article/pii/000187089190003P.

[10] Soumya Kundu, Karanjit Kalsi, and Scott Backhaus. Approximating flexibility in distributed energy resources: A geometric approach. In *2018 Power Systems Computation Conference (PSCC)*, pages 1–7. IEEE, 2018.

[11] Prabha Kundur. *Power System Stability and Control*. McGraw-Hill, 1994.

[12] A. P. Leite, C. L. T. Borges, and D. M. Falcao. Probabilistic wind farms generation model for reliability studies applied to brazilian sites. *IEEE Transactions on Power Systems*, 21(4):1493–1501, Nov 2006. ISSN 0885-8950. DOI: 10.1109/TPWRS.2006.881160.

[13] Tom Mens and Pieter Van Gorp. A taxonomy of model transformation. *Electronic Notes in Theoretical Computer Science*, 152:125–142, 2006.

[14] James R Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

[15] Farrokh Rahimi and Ali Ipakchi. Demand response as a market resource under the smart grid paradigm. *IEEE Transactions on smart grid*, 1(1):82–88, 2010.

[16] Lorenzo Reyes-Chamorro, Andrey Bernstein, Jean-Yves Le Boudec, and Mario Paolone. A composable method for real-time control of active dis-

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

tribution networks with explicit power setpoints. part II: Implementation and validation. *Electric Power Systems Research*, 125:265 – 280, 2015. ISSN 0378-7796. DOI: https://doi.org/10.1016/j.epsr.2015.03.022. URL http://www.sciencedirect.com/science/article/pii/S0378779615000899.

[17] Emily Riehl. *Category Theory in Context*. Dover Publications, 2016. URL http://www.math.jhu.edu/~eriehl/context.pdf.

[18] Mark F. Ruth, Monte S. Lunacek, and Birk Jones. Impacts of using distributed energy resources to reduce peak loads in vermont. 11 2017. DOI: 10.2172/1411137.

[19] Patrick Schultz and Ryan Wisnesky. Algebraic data integration. *Journal of Functional Programming*, 27:e24, 2017. DOI: 10.1017/S0956796817000168.

[20] Patrick Schultz, David I. Spivak, Christina Vasilakopoulou, and Ryan Wisnesky. Algebraic Databases. *arXiv e-prints*, art. arXiv:1602.03501, Feb 2016.

[21] P. Selinger. *A Survey of Graphical Languages for Monoidal Categories*, pages 289–355. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-12821-9. DOI: 10.1007/978-3-642-12821-9_4. URL https://doi.org/10.1007/978-3-642-12821-9_4.

[22] David I. Spivak. Functorial data migration. *Information and Computation*, 217:31 – 51, 2012. ISSN 0890-5401. DOI: https://doi.org/10.1016/j.ic.2012.05.001. URL http://www.sciencedirect.com/science/article/pii/S0890540112001010.

[23] Michael Stay and L. G. Meredith. Representing operational semantics with enriched Lawvere theories. *arXiv e-prints*, art. arXiv:1704.03080, Apr 2017.

[24] Perdita Stevens. Generative and transformational techniques in software engineering II. chapter A Landscape of Bidirectional Model Transformations, pages 408–424. Springer-Verlag, Berlin, Heidelberg, 2008. ISBN 978-3-540-88642-6. DOI: 10.1007/978-3-540-88643-3_10. URL http://dx.doi.org/10.1007/978-3-540-88643-3_10.

[25] Frank Trollmann and Sahin Albayrak. Extending model to model transformation results from triple graph grammars to multiple models. In Dimitris Kolovos and Manuel Wimmer, editors, *Theory and Practice of Model Transformations*, pages 214–229, Cham, 2015. Springer International Publishing. ISBN 978-3-319-21155-8.

[26] W. Tushar, S. Huang, C. Yuen, J. A. Zhang, and D. B. Smith. Synthetic gen-

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

eration of solar states for smart grid: A multiple segment markov chain approach. In *IEEE PES Innovative Smart Grid Technologies, Europe*, pages 1–6, Oct 2014. DOI: 10.1109/ISGTEurope.2014.7028832.

[27] Ray Daniel Zimmerman, Carlos Edmundo Murillo-Sánchez, and Robert John Thomas. Matpower: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 26(1):12–19, Feb 2011. ISSN 0885-8950. DOI: 10.1109/TPWRS.2010.2051168.

Anand, Dhananjay; Pollard, Blake; Breiner, Spencer; Nolan, John; Subrahmanian, Eswaran.
"Compositional models for power systems."
Paper presented at ACT 2019 - Applied Category Theory Conference, Oxford, United Kingdom. July 15, 2019 - July 19, 2019.

Extended Abstracts of The Second Pacific Rim Thermal Engineering Conference
December 13-17, 2019, Maui, Hawaii, USA

PRTEC-24081

RADNNET-MBL: A NEURAL NETWORK APPROACH FOR EVALUATION OF ABSORPTIVITY AND EMISSIVITY OF NON-GRAY COMBUSTION GAS MIXTURE BETWEEN FINITE AREAS AND VOLUMES

Walter W. Yuen[1*], Wai Cheong Tam[2]
[1] Department of Mechanical Engineering, Santa Clara University, USA
[2] Fire Research Division, National Institute of Standards and Technology, USA

ABSTRACT

RADNNET-MBL (RADiation-Neural-NETwork with Mean-Beam-Length) is developed to provide a computationally efficient and accurate method for the evaluation of radiative heat transfer between arbitrary rectangular surfaces in a three-dimensional enclosure with a non-gray combustion medium. Exact mean beam lengths between an infinitesimal area and a finite rectangular area are calculated for a wide range of geometric configurations and optical thicknesses. For a specific geometric configuration, the effect of optical thickness on mean beam lengths is examined. Based on numerical experiment, a constant average mean beam length is defined and shown to be effective in generating accurate prediction of transmissivity over all optical thicknesses. Utilizing the averaged mean beam lengths together with RADNNET (a neural network-based correlation), exchange factors between two finite rectangular areas with an intervening non-gray combustion mixture can readily be obtained. A case study is presented. Results show that RADNNET-MBL provides promising accuracy (with absolute error less than 1%) with a significant reduction in computational effort.

KEYWORDS: Mean beam length, neural network, non-gray, three-dimensional, radiation heat transfer

1. INTRODUCTION

The evaluation of radiative heat transfer with the presence of a participating medium consisting of combustion products (i.e., H_2O, CO_2, and soot particulate) in a three-dimensional enclosure is numerically complex. Even for a one-dimensional isothermal homogeneous medium, the absorptivity is a complicated function of six independent variables, including source temperature, mixture temperature, soot volume fraction, and optical thicknesses of H_2O, CO_2, and CO [1]. In order to account for the spectral behavior of the gaseous mixture, a direct numerical integration using realistic spectral data is required. To account for the geometric effect between arbitrary surfaces and the medium, another direct numerical integration is also needed. Previous work [2] shows that for a three-dimensional rectangular enclosure, the determination of necessary exchange factors will require 480 million numerical evaluations for each time-step in a typical engineering calculation. In the area of fire protection engineering, a common practice to overcome this numerical bottleneck is to utilize simplified methods [3,4]. These simplified methods have never been benchmarked with exact solutions and their accuracy is therefore highly uncertain. Recent studies [5,6] have shown that the use of such simplified methods can potentially lead to substantial errors (higher than 100 % or up to 1000 %). Indeed, the lack of a mathematically validated and computational efficient methodology which would allow non-radiation experts to implement the correct physics of radiative transfer into practical engineering design calculations is a serious obstacle, particularly to the fire protection community, in understanding the effect of radiative heat transfer for fire safety consideration.

In previous works, using realistic spectral data of RADCAL [7] (a narrow band model), a neural network based model, RADNNET, was developed [1] and shown to be a computational efficient approach to determine the total absorption characteristics of a one-dimensional non-gray combustion gas mixture. RADNNET was subsequently extended to RADNNET-ZM [8], which evaluates total emissivity and absorptivity of a

*Corresponding Author: wwyuen@scu.edu

Tam, Wai Cheong; Yuen, Walter.
"RADNNET-MBL: A Neural Network Approach for Evaluation of Absorptivity and Emissivity of Non-Gray Combustion Gas Mixture Between Finite Areas and Volumes."
Paper presented at Second Pacific Rim Thermal Engineering Conference, Maui, HI, United States. December 13, 2019 - December 17, 2019.

homogeneous combustion medium in a three-dimensional rectangular enclosure using direct numerical integrations. The objective of this paper is to show that by utilizing the concept of mean beam length (MBL), the computational efficiency for the evaluation of the total emissivity and absorptivity of a three-dimensional non-gray combustion gas mixture can be significantly improved. In the following sections, the mathematical formulation of MBL, for two fundamental rectangular geometries, is first presented. While MBL is generally a function of optical thickness, numerical data will be generated to show that an average MBL can be defined and used to correlate the absorption effect over all optical thicknesses to a high degree of accuracy. Using superposition and the tabulated average MBL for the two fundamental rectangular geometric configurations, RADNNET-MBL is developed. Numerical results are generated to demonstrate both the accuracy and numerical efficiency of RADNNET-MBL.

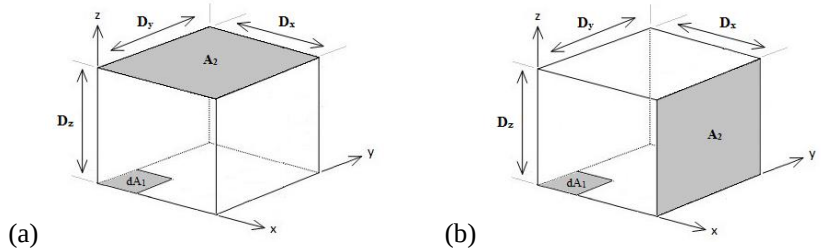


Fig 1. Exchange factor configuration for a) parallel areas and b) perpendicular areas.

2. MATHEMATICAL FORMULATION

Consider a one-zone enclosure filled with a homogenous mixture of water vapor, carbon dioxide, and soot particulate with arbitrary dimensions of D_x, D_y, and D_z as shown in Figs. 1a and 1b, the analysis of radiative heat transfer to the bounding surfaces requires the evaluation of surface-surface exchange factors. Specifically, for perpendicular areas as shown in Fig. 1b (the mathematical development for the geometry shown in Fig. 1a is similar and will not be presented in this paper due to page limitation, the derivation will be available to the reader upon request), the radiative heat transfer from a differential area dA_1 to area A_2, Q_{d1-2}, is given by:

$$Q_{d1-2} = \frac{W_{d1} dA_1}{\pi} \iint \frac{e^{-kr_{D_x D_z}}}{r^4} dA_2 = W_{d1} ds_1 s_2 \tag{1}$$

where W_{d1} is the radiosity leaving the differential area dA_1, r is the center-to-center distance between the differential area and the finite area A_2, and k is the local absorption coefficient of the mixture. From Eq. (1), the exchange factor can be rewritten as:

$$ds_1 s_2 = \frac{dA_1}{\pi} \iint \frac{e^{-kr_{D_x D_z}}}{r^4} dA_2 = dA_1 F_{d1-2} \tau_{d1-2} \tag{2}$$

with

$$F_{d1-2} \tau_{d1-2} = \iint \frac{e^{-kr_{D_x D_z}}}{\pi r^4} dA_2 \tag{3}$$

and $r = \left(D_z^2 + D_y^2 + D_x^2\right)^{\frac{1}{2}}$ and $dA_2 = dzdy$. Let $s = \left(D_z^2 + D_y^2\right)^{\frac{1}{2}}$ and $dA_2 = sdsd\phi$ with $D_z = scos\phi$, Eq. (3) can be written in polor coordinates as:

$$F_{d1-2} \tau_{d1-2} = \int_0^{S_{max}} \int_{\phi_{min}}^{\phi_{max}} \frac{e^{-kr_{D_x scos\phi}}}{\pi r^4} sdsd\phi \tag{4}$$

Eq. (4) can be further integrated in the angular direction for a different range of variables, leading to three one-dimensional integrals as follows:

Tam, Wai Cheong; Yuen, Walter.
"RADNNET-MBL: A Neural Network Approach for Evaluation of Absorptivity and Emissivity of Non-Gray Combustion Gas Mixture Between Finite Areas and Volumes."
Paper presented at Second Pacific Rim Thermal Engineering Conference, Maui, HI, United States. December 13, 2019 - December 17, 2019.

$$F_{d1-2}\tau_{d1-2} = g_1 + g_2 + g_3 \tag{5}$$

with

$$g_1 = \int_0^{\frac{D_y}{D_x}} \frac{e^{-kD_x\sqrt{1+\eta^2}}}{\pi(1+\eta^2)^2}\eta^2 d\eta \tag{6a}$$

$$g_2 = \frac{D_y}{\pi D_x}\int_{D_y/D_x}^{\sqrt{\frac{D_y^2/}{D_x^2}+\frac{D_z^2}{D_x^2}}} \frac{e^{-kD_x\sqrt{1+\eta^2}}}{(1+\eta^2)^2}\eta d\eta \tag{6b}$$

$$g_3 = -\frac{1}{\pi}\int_{\frac{D_z}{D_x}}^{\sqrt{\frac{D_y^2/}{D_x^2}+\frac{D_z^2}{D_x^2}}} \frac{e^{-kD_x\sqrt{1+\eta^2}}}{(1+\eta^2)^2}\sqrt{\eta^2-\frac{D_z^2}{D_x^2}}\eta d\eta \tag{6c}$$

and $\eta = s/D_x$. Note that F_{d1-2} is the view factor between dA_1 and A_2 given by:

$$F_{d1-2} = \frac{1}{2\pi}\left[tan^{-1}\left(\frac{D_y}{D_x}\right) - \frac{1}{\sqrt{1+\frac{D_z^2}{D_x^2}}}tan^{-1}\left(\frac{D_y}{D_x}/\sqrt{1+\frac{D_z^2}{D_x^2}}\right)\right] \tag{7}$$

Introducing the concept of mean beam length (physically, it is the length of a corresponding one dimensional line-of-sight which yields the same transmissivity):

$$\tau_{d1-2} = \exp(-kL_m) \tag{8}$$

the mathematical expression of mean beam length is:

$$L_m/D_x\left(\frac{D_y}{D_x}, \frac{D_z}{D_x}, kD_x\right) = -ln\left[\frac{(g_1+g_2+g_3)}{F_{d1-2}}\right]/(kD_x) \tag{9}$$

Note that the normalized mean beam length (L_m/D_x) is only a function of the optical thickness kD_x and the dimensionless geometric parameters, (D_y/D_x and D_z/D_x).

Figures 2 show the effect of optical thickness and geometric effect on the mean beam lengths and the exchange factors for the 6 different geometric configurations. It is interesting to note that as the vertical dimension (D_z/D_x) increases, the MBL and the associated exchange factor approach a constant value. Physically, the upper portion of the area has decreasing contribution to the total radiative exchange as the vertical dimension increases. The MBL thus approaches an asymptotic constant value as D_z/D_x increases.

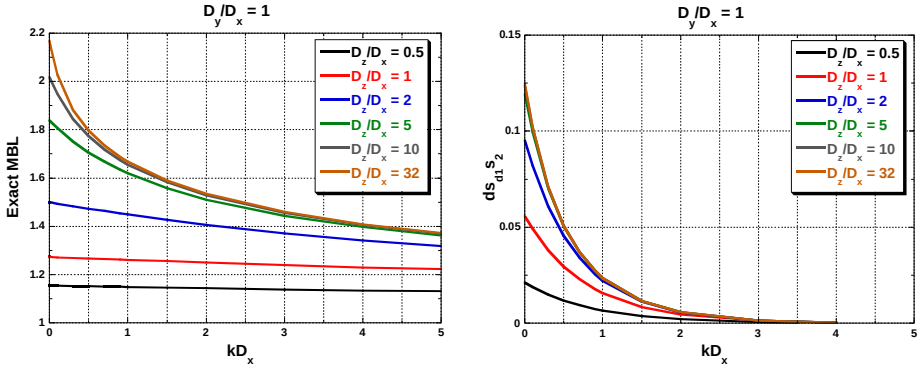


Fig. 2 a) Exact MBLs and b) exchange factors as a function of optical thickness for different D_z/D_x.

Tam, Wai Cheong; Yuen, Walter.
"RADNNET-MBL: A Neural Network Approach for Evaluation of Absorptivity and Emissivity of Non-Gray Combustion Gas Mixture Between Finite Areas and Volumes."
Paper presented at Second Pacific Rim Thermal Engineering Conference, Maui, HI, United States. December 13, 2019 - December 17, 2019.

While the exact value of the MBL varies with optical thickness, numerical results show further that the exchange factor can be correlated accurately with a constant average MBL. Mathematically, this average MBL is defined by first introducing an accumulative error function for a general dimensionless variable η as follows:

$$E(\eta) = \sum_{i=1}^{N} \left(\tau_{d1-2} - e^{-(kD_x)_i \eta}\right)^2 \qquad (10)$$

where N is the total number of optical thicknesses being considered. The average MBL is defined to be the value of η at which the accumulative error function is a minimum. Mathematically, it is given by:

$$\frac{dE}{d\eta}\left(\frac{L_{m,a}}{D_x}\right) = -\sum_{i=1}^{N} 2kD_x e^{-(kD_x)_i\left(\frac{L_{m,a}}{D_x}\right)} \left(\tau_{d1-2} - e^{-(kD_x)_i\left(\frac{L_{m,a}}{D_x}\right)}\right) = 0 \qquad (11)$$

The average mean beam lengths, the exact exchange factors, and the approximated exchange factors for the 6 cases are listed in Table 1. It can be seen that a constant average MBL is an excellent approximation to the radiative heat transfer for different geometry and optical thicknesses.

Table 1 Relative difference in between exact and approximated exchange factor for different geometry and optical thicknesses with D_y/D_x = 1.0.

kD_x	D_z/D_x	Average MBL	Exchange Factor (Approximated)	Exchange Factor (Exact)	Relative Difference
0.1	1	1.2598	0.04914	0.04908	-0.00006
0.1	5	1.6411	0.10095	0.09930	-0.00165
0.1	10	1.6946	0.10419	0.10159	-0.00260
1.0	1	1.2598	0.01581	0.01578	-0.00003
1.0	5	1.6411	0.02305	0.02356	0.00051
1.0	10	1.6946	0.02267	0.02357	0.0009

Utilizing the concept of average MBL, the radiative heat transfer from a black differential area dA_1 with temperature T_w to a finite area A_2 as shown in Figs 1a and 1b, with an intervening combustion gas mixture, can now be written as:

$$Q_{d1-2} = \sigma T_w^4 ds_1 s_2 = dA_1 F_{d1-2} \int_0^\infty e_{\lambda,b}(T_w)\tau_{d1-2,\lambda} d\lambda = \sigma T_w^4 dA_1 F_{d1-2} \int_0^\infty e^{-k_\lambda L_{m,a}} d\lambda \qquad (12)$$

Since the average MBL is constant for a given geometry, the spectral integral in Eq. (2) can be readily evaluated using RADNNET. For a given source temperature, T_w, mixture temperature, T_g, absorbing gas partial pressure, P_g, mole fraction of CO_2, X_{CO_2}, and soot volume fraction, f_v, the differential exchange factor can be written as:

$$ds_1 s_2 = dA_1 F_{d1-2}\left[1 - \alpha_{d1-2}\left(T_g, T_w, P_g L_{m,a}, \chi_{co_2}, f_v L_{m,a}\right)\right] \qquad (13)$$

where α is the total absorptivity evaluated based on RADNNET. It is important to note that for a rectangle with arbitrary dimensions and location relative to the differential area dA_1, either in the parallel or perpendicular orientation, the exchange factors between two finite areas can be generated by a finite sum or differences of rectangle with the geometry as shown in Figs. 1a and 1b. The differential exchange factor between dA_1 and a rectangle of arbitrary dimension (either in the parallel or perpendicular orientation) can thus be written as a finite sum or differences with terms similar to Eq. (13). This is a signification reduction in computational time and effort compared to direct numerical integration. For the radiative exchange between two finite rectangular areas A_1 and A_2 (either in the parallel or perpendicular orientation), the exchange factor can be generated by a single numerical integration. This is the basis of RADNNET-MBL.

Tam, Wai Cheong; Yuen, Walter.
"RADNNET-MBL: A Neural Network Approach for Evaluation of Absorptivity and Emissivity of Non-Gray Combustion Gas Mixture Between Finite Areas and Volumes."
Paper presented at Second Pacific Rim Thermal Engineering Conference, Maui, HI, United States. December 13, 2019 - December 17, 2019.

3. RESULTS AND DISCUSSION

A verification case is provided to validate the accuracy and to examine the numerical efficiency of RADNNET-MBL for the evaluation of exchange factor for two perpendicular surfaces in a 1 m^3 cubical enclosure. The surfaces are black. The temperature of the surfaces and the mixture temperatures are uniform and maintained at 1000 K. The total pressure of the gas mixture is kept at 101.325 kPa and consists of water vapor and soot particulates.

Results with differential absorption gas pressure and soot volume fraction are generated and they are summarized in Table 2. It can be seen that the relative difference associated with the exact and the RADNNET-MBL generated surface-surface exchange factors is less than 1%. RADNNET-MBL, however, speeds up the evaluation dramatically.

Table 2 Summary of results for the verification cases.

Case	$T_g = T_w$ (K)	P_g	f_v	S_1S_2	CPU (s)	S_1S_2(MBL)	CPU (s)
1	1000	0	0	0.1999	1.562E-2	0.1999	1.562E-2
2	1000	0	5E-8	0.1901	3.125E-2	0.1908	1.562E-2
3	1000	30	0	0.1615	27.70	0.1630	7.813E-2
4	1000	30	5E-8	0.1539	6.547	0.1556	9.375E-2

4. CONCLUSIONS

RADNNET-MBL is presented. Based on numerical study, it is demonstrated that the new approach is computationally efficient and accurate for the evaluation of radiative heat transfer between arbitrary rectangular surfaces in a three-dimensional enclosure with a non-gray combustion medium. For four test cases, the relative difference associated with the surface-surface exchange factor is shown to be less than 1%. Using RADNNET-MBL, the evaluation of the exchange factor for some cases can be reduced by more than a factor of 100.

REFERENCES

[1] Yuen, W. W., "RAD-NNET, a Neural Network Based Correlation Developed for a Realistic Simulation of the Non-gray Radiative Heat Transfer Effect in Three-dimensional Gas-particle Mixtures." *International Journal of Heat and Mass Transfer*, 52, 3159–3168 (2009).

[2] Tam, W. C., *Analysis of Heat Transfer in a Building Structure Accounting for the Realistic Effect of Thermal Radiation Heat Transfer*, Ph.D. Thesis, The Hong Kong Polytechnic University (2013).

[3] Hurley, M. J., Gottuk, D. T., Hall Jr, J. R., Harada, K., Kuligowski, E. D., Puchovsky, M., ... & Wieczorek, C. J., *SFPE Handbook of Fire Protection Engineering*, New York: Springer, pp. 102–137 (2016).

[4] Edwards, D. K., in *Handbook of Heat Transfer Fundamentals*, New York: McGraw-Hill, (1985).

[5] Yuen, W. W., Tam, W. C., Chow, W. K., "Assessment of Radiative Heat Transfer Characteristics of a Combustion Mixture in a Three-dimensional Enclosure using RAD-NETT (with Application to a Fire Resistance Test Furnace)." *International Journal of Heat and Mass Transfer*, 68, pp. 383–390 (2014).

[6] Tam, W. C., Yuen, W. W., "Assessment of Radiation Solvers for Fire Simulation Models Using RADNNET-ZM," *11th AOSFST*, (2018).

[7] Grosshandler, W. L., "RADCAL: a Narrow-band Model for Radiation." *Calculations in a Combustion Environment, NIST Technical Note 1402* (1993).

[8] Tam, W. C., Yuen, W. W., "OpenSC – an Open-source Calculation Tool for Combustion Mixture Emissivity/Absorptivity." *NIST Technical Note* (2019).

Tam, Wai Cheong; Yuen, Walter.
"RADNNET-MBL: A Neural Network Approach for Evaluation of Absorptivity and Emissivity of Non-Gray Combustion Gas Mixture Between Finite Areas and Volumes."
Paper presented at Second Pacific Rim Thermal Engineering Conference, Maui, HI, United States. December 13, 2019 - December 17, 2019.

Title: Measuring Water Flow Rate in a Flexible Fire Hose using an Accelerometer

Authors: Christopher U. Brown, Gregory W. Vogl, and Wai Cheong Tam

National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, MD 20899, 301-975-5852, christopher.brown@nist.gov, gregory.vogl@nist.gov, waicheong.tam@nist.gov

Abstract

A wired sensor network was created to measure water-flow rate in a fire hose. An integrated electronic piezoelectric (IEPE) accelerometer was chosen as the sensor to measure the flow rate based on the vibrations generated by water flowing through a fire hose. These sensors are small, lightweight, and they can attach to the outside of the hose, not obstructing the water's flow path. A relationship was determined between the flow rate of the water and vibration detected by the accelerometer for a range of flow rates. The raw acceleration signal was used to calculate two metrics: the dominant frequency and the standard deviation of acceleration. In a future study, the relationship between the dominant-frequency metric and the flow rate will be applied to a wireless accelerometer network. The relationship will be used to determine real-time, fire hose, flow rate critical for improving situational awareness on the fireground.

Keywords: Accelerometers, fire hose, flow induced vibration, flow rate, hose vibration, dominant frequency, smart firefighting, sensors, water flow rate measurement, wired sensor network.

Introduction

Placing and flowing water through initial, intermediate, and final hose lines is very important for the success of a fire attack. The water discharged from the hose nozzle cools the environment, which improves the survival of trapped occupants, protects the firefighters from excessive heat, and extinguishes the fire. Therefore, hoses are simultaneously a firefighter's and occupant's lifelines [1]. Knowing the water flow rate through a fire hose is a critical part of fire suppression and situational awareness, especially for zero flow conditions, which are not uncommon.

Applying 'smart' technology to a fire hose could improve 1) the awareness of the hose's current status and 2) the chance of a successful fire attack. Harnessing the power of 'smart' technology to improve situational awareness of the hoses was a part of the vision of Smart Fire Fighting as documented in the Research Roadmap for Smart Fire Fighting [2]. A 'smart' system uses sensors to collect data, provides the data in an understandable format to a user, then allows the user to make decisions. Today, the users are humans, but tomorrow they will include software.

Human firefighters are the backbone of the fire service. The safety of firefighters who risk their lives on the fireground could benefit from a smart sensor system that could perform two tasks: determine if water is flowing at the fire-hose nozzle and communicate this information back to a human controller at the fire engine.

Brown, Christopher; Vogl, Gregory; Tam, Wai Cheong.
"Measuring Water Flow Rate in a Flexible Fire Hose using an Accelerometer."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

Currently, communication between the firefighter at the nozzle and the pump operator or incident commander (IC) is typically done using radios. The firefighter at the nozzle, or his backup, should be able to communicate by radio with the pump operator or IC to provide feedback about water flow. However, this is not always possible due to competing radio traffic and performance of tasks that require two hands such as advancing the hose and conducting suppression activities.

Presently, water pressure measured at the fire engine's pump panel is used by the pump operator to determine if water is flowing at the hose nozzle. Fireground threats to normal water flow such as hose damage and hose blockage can make reliable decisions on water pressure misleading. A pressure loss indicated on the fire engine's pump panel may occur when water flows from the nozzle as intended, or unintentionally through a ruptured hose. A ruptured hose can occur as a result of wear and tear, a burn hole in the hose, a leaking coupling, or from being crushed under a vehicle tire or structure debris.

Sufficient water pressure may show at the pump panel even if the hose is partially or fully blocked preventing water, or allowing too little water, from reaching the nozzle. A charged hose line advanced inside a structure could become partially blocked as a result of being crimped around a sharp corner or past a piece of furniture. A hose pinched under a door, under a piece of furniture, under a vehicle tire, or under fallen debris could also reduce water flow at the nozzle. Water flow through the hose could be fully blocked by a closed in-line valve, debris in the hose, or by a closed nozzle bale that cannot be opened by an incapacitated firefighter.

A reliable way for the IC to know that water is flowing from the hose nozzle is to have real-time water flow information sent to them. The goal of this study is to provide that information digitally by developing a wired sensor network to measure water flow in a fire hose. Our approach, which is detailed below, is to collect vibration data, convert it into a flow rate using a well-known algorithm, then display the flow rate at the incident command post.

Methods

The commercial fire-attack hose used in this study has a nominal 4.5 cm (1.75 in) inner diameter and was approximately 15 m (50 ft) long. The initial flow rate was set in 19 LPM (5 GPM) increments between approximately 0 LPM and 606 LPM (160 GPM). For consistency, the abbreviation 'LPM' represents L/min and 'GPM' represents gallons/min. A commercial turbine flow meter was used at the nozzle to measure the reference flow rate during the study. For a variety of reasons, the reference flow rate drifted approximately ±3.8 LPM (1.0 GPM) at the highest reference flow rates but drifted less than approximately ±1.9 LPM (0.5 GPM) at the lower reference flow rates.

Several different types of flow meters were assessed for use in this study such as: turbine, electromagnetic, and pressure differential meters. However, these types of meters were typically too large, heavy, and bulky for use with the fire hose. Alternatively, a small, lightweight, exterior accelerometer was chosen for the fire hose application. The four piezoelectric accelerometer sensors (PCB model 288D01 and 352C33) communicated over the wired network to the data acquisition system. The sensors, which stand about 2 cm high, were mounted on bases epoxied to

the outer surface of the hose (Fig. 1). The four accelerometers were positioned along the hose (Fig. 2) [3]. Vibration data from all four accelerometers was collected for 50 consecutive, 3-second-long test intervals at a sampling frequency of 5 kHz.

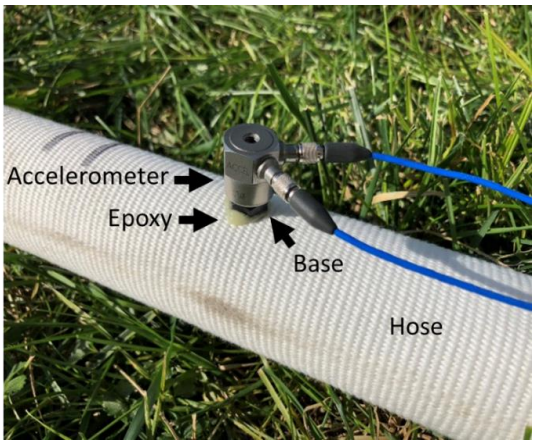


Fig. 1. The accelerometer attached to a base that was epoxied to the exterior fabric of the fire hose at the downstream (Front) location.

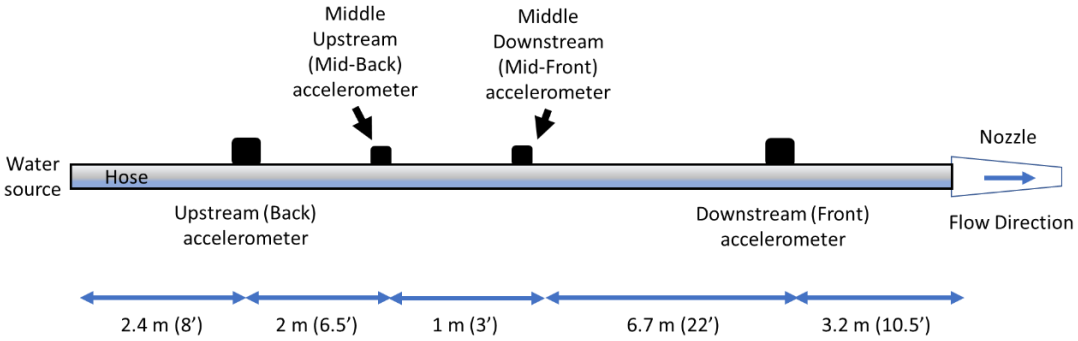


Fig. 2. The four accelerometer locations along the approximately 15 m (50 ft) long fire hose.

Impact testing was done to understand the flexible hose dynamics as well as the dominant frequency of the hose system. An impact hammer was used on the downstream (Front) accelerometer during the following approximate flow rates to determine dominant frequencies: 606 LPM (160 GPM), 454 LPM (120 GPM), 303 LPM (80 GPM), 151 LPM (40 GPM), and 0 LPM. The impact profiles of thirty impacts were collected.

Results and Discussion

The standard deviation of acceleration was calculated from the raw acceleration data and plotted versus flow rate at each of the four accelerometer locations along the hose (Fig. 3). The standard deviation of acceleration generally increased until peaking and then decreased at all four accelerometer locations. The standard deviation values at the front accelerometer location were larger than the values from the other three locations.

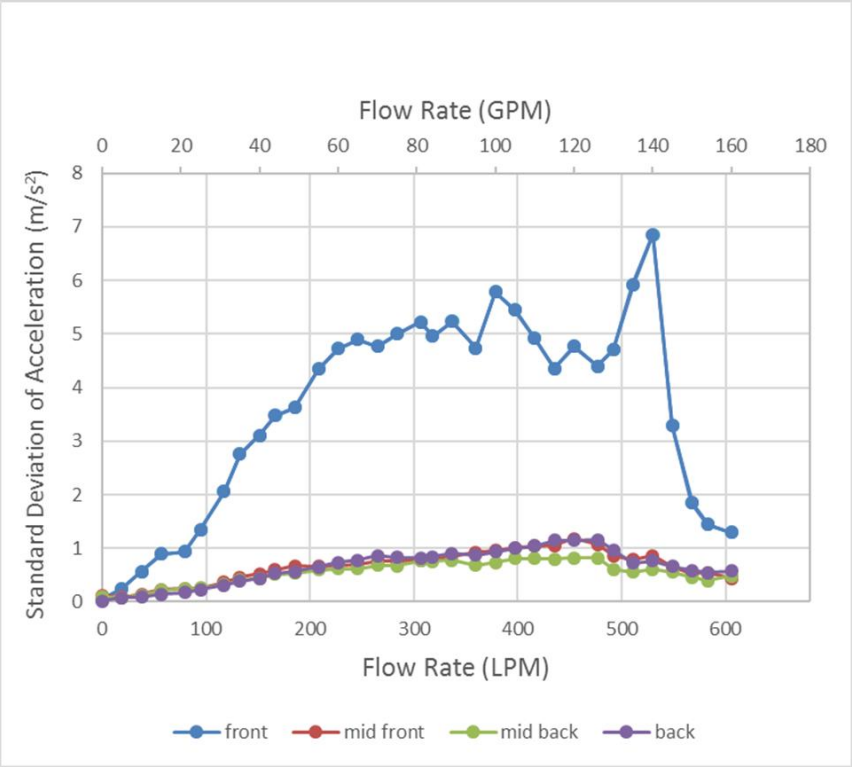


Fig. 3. Typical results of the standard deviation of acceleration versus flow rate for the four accelerometers. Data among the 50 consecutive runs was similar and therefore for visualization purposes only one run was used to represent all the data in the figure.

Based on previous studies using relatively rigid pipe [4-9], a decreasing trend was not expected at the higher flow rates for the flexible hose. The structural dynamic differences between the rigid pipes and the flexible hose could be causing the decreasing trend at higher flow rates although additional research is needed for confirmation. The bell-shaped curve determined at each accelerometer location excludes the standard deviation of acceleration as a metric for determining flow rate over the entire flow range because of the lack of monotonicity; at a single level of standard deviation of acceleration, there are two corresponding flow rates.

The time-domain acceleration data was converted to frequency-domain using a Fast Fourier Transform (FFT). A dominant frequency at each flow rate was observed. A decreasing trend was observed for the front accelerometer only (Fig. 4). When water is flowing through a rigid pipe, the dominant frequency typically does decrease with increasing flow rate [4, 10, 11]. Thus, a similar trend at the front accelerometer was expected for a flexible hose.

To test that hypothesis, we compared the results from two different testing approaches: impact testing and flow-rate testing. The dominant-frequency results as calculated from the impact testing at 5 flow rates were very similar to the results from the flow tests (Fig. 5). This comparison confirmed the determination of the dominant frequency using only the flow tests, and no additional impact testing was needed. What was needed, however, was the ability to

determine the real-time flow rate using dominant frequency as a metric. In this paper we show that it was possible to make that determination using only the data from the accelerometer located closest to the hose nozzle.

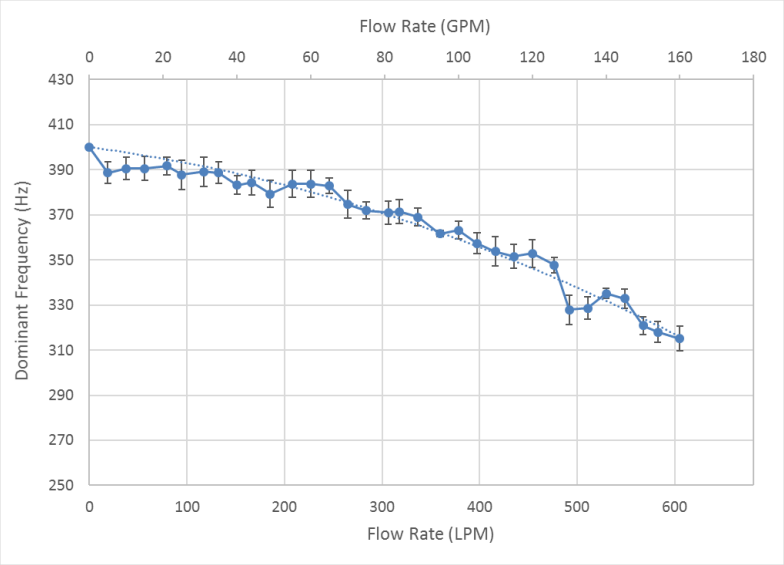


Fig. 4. The mean and standard deviation for the dominant frequency versus flow rate for the Front accelerometer from the flow-rate testing.

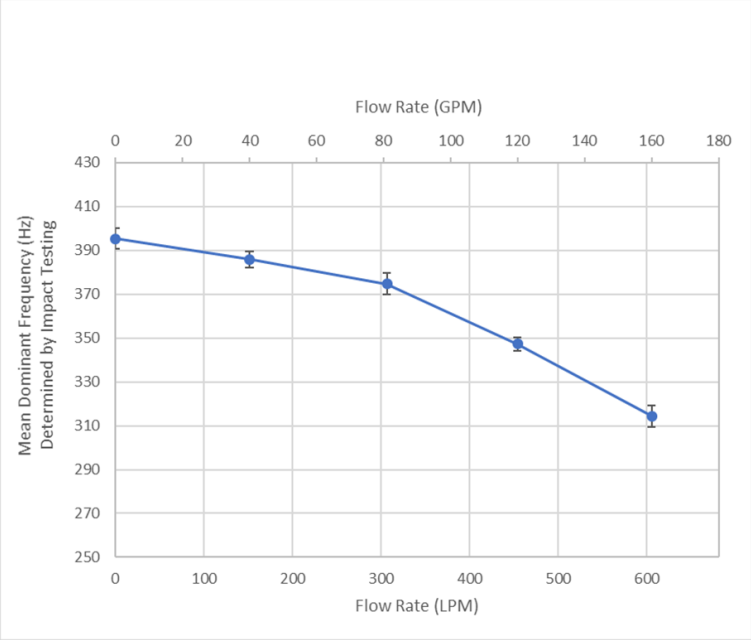


Fig. 5. The mean and standard deviation for the dominant frequency versus flow rate for the Front accelerometer from the impact testing.

Brown, Christopher; Vogl, Gregory; Tam, Wai Cheong.
"Measuring Water Flow Rate in a Flexible Fire Hose using an Accelerometer."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

Conclusion

The goal of this study was to develop the wired accelerometer as a sensor to determine flow rate near the nozzle of a fire hose. The standard deviation of acceleration and dominant frequency were examined to determine which metric would correlate well with flow rate over the entire flow rate range. The dominant frequency metric yielded a somewhat linear, and generally monotonic, relationship with flow rate for an accelerometer located close to the hose nozzle that can be applied for further study. The next step will be to develop the wired accelerometer system into a wireless sensor network and apply a metric based on dominant frequency and other measures to robustly determine flow rate in a fire hose [12]. The final step will be to use the wireless accelerometer network and finalized metric to determine real-time flow in a fire hose to improve fireground situational awareness.

Acknowledgement and Disclaimer

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology (NIST), nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose. The opinions, recommendations, findings, and conclusions in this publication do not necessarily reflect the views or policies NIST or the United States Government.

References

[1] Petrillo AM (2019) Hoses: Large and Small, from LDH to Forestry Lines. *Fire Apparatus and Emergency Equipment*. June, 24(6). https://digital.fireapparatusmagazine.com/fireapparatus/june_2019/MobilePagedReplica.action?pm=2&folio=16#pg16

[2] Hamins A, Grant C, Bryner N, Jones A, Koepke G (2015) Research Roadmap for Smart Fire Fighting, National Institute of Standards and Technology, Gaithersburg, MD, NIST Special Publication (SP) 1191. https://doi.org/10.6028/NIST.SP.1191

[3] Brown CU, Vogl GW, Tam WC (2019) Measuring Water Flow Rate for a Fire Hose Using Wired Accelerometers for Smart Fire Fighting. National Institute of Standards and Technology, Gaithersburg, MD, NIST Technical Note (TN) (in draft).

[4] Evans RP, Blotter JD, Stephens AG (2004) Flow rate measurements using flow-induced pipe vibration. *Transactions of the ASME* 126: 280-285. DOI: 10.1115/1.1667882

[5] Pittard MT, Evans RP, Maynes RD, and Blotter JD (2004) Experimental and numerical investigation of turbulent flow induced pipe vibration in fully developed flow. *Review of Science Instruments* 75(7): 2393-2401. Doi: 10.1063/1.1763256

[6] Thompson AS, Maynes D, and Blotter JD (2010) Internal turbulent flow induced pipe vibrations with and without baffle plates, *Proceedings of the ASME 2010 3rd Joint US-European Fluids Engineering Summer Meeting and 8th International Conference on Nanochannels, Microchannels, and Minichannels*. Aug 1-5, Montreal, Canada.

[7] Medeiros KAR, Barbosa CRH, and Oliveira EC (2015) Flow measurement of piezoelectric accelerometers: application in the oil industry. *Petroleum Science and Technology* 33:1402-1409. DOI:10.1080/10916466.2015.1044613

[8] Medeiros KAR, Oliveira FLA, Barbosa CRH, and Oliveira EC (2016) Optimization of flow rate measurement using piezoelectric accelerometers: application in water industry. *Measurement* 91:576-581.

[9] Lannes DP, Gama AL, and Bento TFB (2018) Measurement of flow rate using straight pipes and pipe bends with integrated piezoelectric sensors. *Flow Measurement and Instrumentation* 60: 208-216.

[10] Blevins RD (1977) Flow-Induced Vibration. Van Nostrand Reinhold Company, NY, pp 287-311.

[11] Campagna MM, Dinardo G, Fabbiano L and Vacca G (2015) Fluid flow measurements by means of vibration monitoring. *Meas. Sci. Technol.* 26. DOI:10.1088/0957-0233/26/11/115306

[12] Brown CU, Vogl GW, Tam WC (2019) Measuring Water Flow Rate for a Fire Hose Using a Wireless Sensor Network for Smart Fire Fighting. National Institute of Standards and Technology, Gaithersburg, MD, NIST Technical Note (TN) (in draft).

Brown, Christopher; Vogl, Gregory; Tam, Wai Cheong.
"Measuring Water Flow Rate in a Flexible Fire Hose using an Accelerometer."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

Low-Power Wide Area Networks (LPWAN) for Communications of Mobile Sensor Data

Sebastian Barillaro[*,†]
Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg MD USA
sebastian.barillaro@nist.gov
barillaro@inti.gob.ar

Sokwoo Rhee[†]
Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg MD USA
sokwoo.rhee@nist.gov

Gustavo Escudero[*]
Departamento de validación de dispositivos y sistemas electrónicos
Instituto Nacional de Tecnología Industrial
Buenos Aires, Argentina
gescudero@inti.gob.ar

Raghu Kacker[††]
Information Technology Laboratory
National Institute of Standards and Technology
Gaithersburg MD USA
raghu.kacker@gmail.com

Lee Badger[††]
Information Technology Laboratory
National Institute of Standards and Technology
Gaithersburg MD USA
lee.badger@nist.gov

D. Rick Kuhn[††]
Information Technology Laboratory
National Institute of Standards and Technology
Gaithersburg MD USA
d.kuhn@nist.gov

ABSTRACT

There are multiple options for communication of data to and from mobile sensors. For tracking systems, Global Navigation Satellite System (GNSS) is often used for localization and mobile-phone technologies are used for transmission of data. Low-powerwide area networks (LPWAN) is a newer option for sensor networks including mobile sensors.

We developed a tracking system use case application using LPWAN as communication channel for mobile sensor data. The choice of LPWAN has pros and cons. In this paper, we discuss the differences between LPWAN and other technologies as communication channel for sensor networks. We describe the LPWAN test setup and analyze its characteristics including transmission frequency, coverage, latency and communication range.

* Department of electronic devices and systems validation, National Institute of Industrial Technology, Buenos Aires, Argentina
† Smart Grid and Cyber-Physical Systems Program Office, Engineering Laboratory, National Institute of Standards and Technology, U.S. Department of Commerce
†† Information Technology Laboratory, National Institute of Standards and Technology, U.S. Department of Commerce

CCS CONCEPTS

Network performance analysis, Network components, Embedded and cyber-physical systems, Sensor networks

KEYWORDS

IoT, Internet of Things, LPWAN, Tracker System, LoRa, LoRaWAN, Sensor Network

ACM Reference format:

Sebastian Barillaro, Sokwoo Rhee, Gustavo Escudero, Raghu Kacker, Mark Lee Badger, D. Rick Kuhn. 2019. Low-Power Wide Area Networks (LPWAN) for Communications of Mobile Sensor Data. In *Proceedings of the 2nd ACM/EIGSCC Symposium On Smart Cities and Communities (SCC 2019). ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1234567890*

1 Introduction

The emergence of low power wireless technologies has catalyzed broad adoption of Internet of Things (IoT) applications over the past several years. A typical IoT implementation includes a number of components and building blocks; one of the fundamental building blocks of IoT implementation is the communication and networking layer. Due to diverse requirements unique to each IoT application, a number of different technologies are being developed and adopted for communication and networking.

For the applications involving wireless sensors and actuators, the selection of wireless communication technology becomes a critical task. There are a number of fundamental characteristics to be considered when designing a wireless communication technology for an IoT application, and it should be chosen based on a careful analysis of the application requirements. Most of the wireless communication technologies will come with some type

Barillaro, Sebastian; Rhee, Sokwoo; Kacker, Raghu; Badger, Mark; Kuhn, David; Escudero, Gustavo.
"Low-Power Wide Area Networks (LPWAN) for Communications of Mobile Sensor Data."
Paper presented at 2nd ACM/EIGSCC Symposium On Smart Cities and Communities (SCC 2019), Portland, OR, United States. September 10, 2019 - September 12, 2019.

of tradeoff among different characteristics such as data rate, power consumption, communication distance, latency, cost, reliability, robustness, and security. For example, a technology with higher data rate may need to consume more power. Another technology capable of both high data rate and low power may not be able to communicate in a long range. Because of these tradeoffs, a number of wireless technologies such as WiFi, Bluetooth, and Zigbee are discretely used in various IoT applications depending on their requirements.

Low power wide area network (LPWAN) is a wireless communication technology specifically designed to focus on low power and long-range communication. LPWAN is a promising technology with a great potential to be adopted in a number of smart city applications including metering, localization, transportation, and flood monitoring. However, it is important to understand the practical limitation of its characteristics before designing it in an application. To analyze its characteristics, a LPWAN test network was implemented in the main campus of the National Institute of Standards and Technology (NIST) in Gaithersburg, Maryland, USA, with coverage extended to its vicinity area. The test setup was developed to focus on localizations and tracking application. In this paper, the setup of the test network is described, and the test results are presented.

2 LPWAN General Characteristics

LPWAN is not a single technology. It is a set of several technology options with the shared goal of fulfilling specific sensor network requirements. Commercially known as 0G [1], there are plenty of options to choose from. Although these technologies are always prioritized on wider area coverage and minimal energy consumption over data rate, it is possible to apply them to a variety of uses cases which require different degrees of data rate, communication periodicity, latency, distance, etc. Since these technologies are optimized for energy consumption, some of them cannot handle Internet Protocol (IP) Headers, hence are not IP-compliant.

2.1 LPWAN Options

There are many LPWAN options. The list is broad and its border-line is not clearly defined. It is not the intention of this paper to discuss what is, and what is not an LPWAN option nor make an exhaustive analysis about all of them. However, a non-comprehensive list is given as starting point for readers that would desire to dive deep. This list is divided according to whether a license is required or not to operate in a specific frequency.

2.1.1 Cellular LPWAN Options

- LTE Cat 1
- LTE Cat 0
- EC-GSM-IoT
- LTE Cat M1 (also known as LTE-M)
- LTE Cat M2
- LTE Cat NB1 (also known as NB-IoT)
- LTE Cat NB2

2.1.2 Non-Cellular LPWAN Options

- Sigfox
- IngeNU (formerly RMPA)
- LoRa/LoRaWAN
- LoRa/M.O.S.T.
- LoRa/Symphony Link

Due to NIST's IT security restrictions and policies, four main requirements, among others, were considered when selecting a technology to be tested:

- No proprietary solutions.
- Built-in security in specification.
- Ownership of entire platform.
- No Internet connection required.

All these LPWAN options were analyzed based on available literatures. IngeNU did not meet requirements due to be a proprietary specification. Sigfox, LTE-M, and NB-IoT require Internet connection and cannot be deployed as an isolated network without Internet access. LoRa is a proprietary technology, but LoRaWAN Protocol is an open standard. It does not require Internet connection and its deployment can be done entirely on premises at The NIST Campus as a standalone system. LoRaWAN also has built-in security measures and is broadly adopted. LoRaWAN alternatives (M.O.S.T and Symphony Link) also did not meet requirements due to its nature as a proprietary technology.

After considering all these options, LoRa/LoRaWAN was selected as the LPWAN option to study its characteristics.

2.2 LoRa/LoRaWAN

LoRa (short for Long Range) and LoRaWAN (LoRa Wide Area Network) refer to different meanings and concepts. LoRa is a spread spectrum modulation technique derived from chirp spread spectrum (CSS) technology with an operational frequency in the sub-GHz ISM band. The LoRa-Alliance defines regional specification according to local regulations dictated by local telecommunication authorities. The most widely adopted operating frequencies are the US915 band which includes 64 x 125 kHz width channels and 8 x 500 MHz width channels between 902.3 and 914.9 MHz; and the EU868 band, which includes 8 x 125 kHz channels and one 250 kHz channel between 863 to 870 MHz. Full detailed specification is available on the LoRa-Alliance website [2].

LoRaWAN (formerly known as LoRaMAC) is the specification for Medium Access Control (MAC). The specification [3] is maintained by LoRa-Alliance, a consortium with more than 500 members including companies, research institutions and universities. LoRaWAN is designed to minimize and make efficient use of energy by End-Nodes. It is also designed to communicate bi-directionally, although messages from End-Node to Network (uplink) are prioritized over messages from Network

Barillaro, Sebastian; Rhee, Sokwoo; Kacker, Raghu; Badger, Mark; Kuhn, David; Escudero, Gustavo.
"Low-Power Wide Area Networks (LPWAN) for Communications of Mobile Sensor Data."
Paper presented at 2nd ACM/EIGSCC Symposium On Smart Cities and Communities (SCC 2019), Portland, OR, United States. September 10, 2019 - September 12, 2019.

to End-Nodes (downlink). According to LoRa-Alliance, LoRaWAN supports geo-localization of End-Nodes by the Network Server [4]. However, its feasibility is still being discussed [5].

As shown in Figure 1, there are at least 5 main components in any LoRa/LoRaWAN Topology. They are described in sections 2.2.1 to 2.2.5. Section 2.2.6 describe the dynamic of communications.

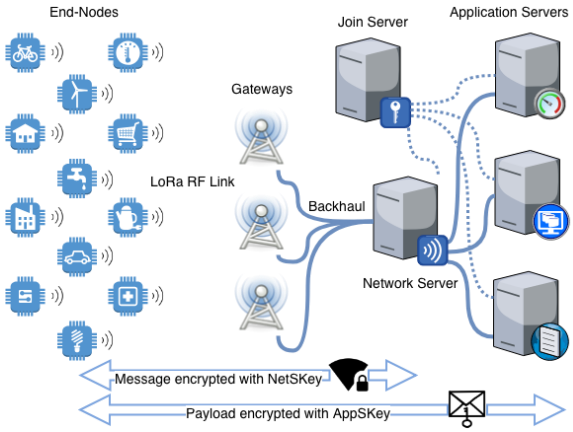


Figure 1: Lora/LoRaWAN Topology

2.2.1 End-node

An End-Node could be any electronic device, which are generally sensors, although they may be actuators as well. An End-Node has a radio module to transmit (up-link) and to receive (down-link) data through LoRa RF Link to/from Cloud. Communication is half-duplex in a specific frequency at a time.

An End-Node first encrypts the payload (usually sensor data) using Application Session Key (AppSKey). Then, it encrypts the entire message, including headers and the encrypted payload with Network Session Key (NetSKey). After these two encryption processes, the End-Node transmits the message in a LoRa packet.

2.2.2 Gateway

The Gateway is an interface between LoRa End-Nodes and the backhaul network. A LoRa interface radio module in a Gateway is used to communicate with End-Nodes through the LoRa RF Link. The LoRa interface radio module in a Gateway is more complex than its counterpart in an End-Node. It is composed of sophisticated and sensitive integrated chips, combining one or more transceivers with at least one Digital Signal Processor (DSP) capable of receiving radio signals in several channels simultaneously. Gateways are usually equipped with a Global Navigation Satellite System (GNSS) sensor which adds timestamps to the messages received.

Gateways do not interpret messages. They do not have any security key to decrypt messages. They only check packet integrity and forward the messages to the Network Server, which is connected to Gateways through a regular IP Network usually implemented over an Ethernet or cellular link. The process is reversed for messages that go from the Network Server to End-Nodes.

2.2.3 Network Server

The Network Server is the manager of network. There is only one Network Server per LoRa Network. It registers End-Nodes, Gateways, and Application Servers. The Network Server receives LoRa frames wrapped in IP packets from gateways, and sort out and discard possible duplicate messages that may be received by more than one gateway connected to the same network. It decrypts LoRa frame using NetSKey and passes the encrypted payload to the respective Application Server. When the network traffic flows in the other direction, the Network Server encrypts messages, including headers and payloads (which are already encrypted with AppSKey, by Application Servers) using NetSKey and assigns them to a gateway to be forwarded to an End-Node. Because AppSKey and NetSKey are different, the Network Server does not have access to the encrypted payloads.

2.2.4 Application Server

Application Servers (there may be more than one per network) are the final destination of encrypted payloads from End-Nodes in the LoRa system. It is worth noting that the payload transmitted and received by the LoRa system may continue its way through the cloud to a final destination where it becomes useful, such as a database or dashboard. However, that part of the IoT System is out of scope of the LoRa/LoRaWAN architecture although it can be conceptually aggregated in an Application Server. Application Servers have the AppSKey to decrypt payloads.

2.2.5 Join Server

Join Server plays a role at the beginning of communications which handles Join-Request from End-Nodes during activation stage. Join Server was added in LoRaWAN Specification [3] v1.1 to support the segregation of key management. A deep analysis of the modifications made in v1.1 can be found in [6].

There might be more components. For example, LoRaWAN v1.1 includes Home Server, Forwarding Server and Serving Server to support Roaming between networks [7].

2.2.6 Protocol Operation

LoRa/LoRaWAN protocol operation may be divide in 2 stages: Activation and Communication. Activation is a process to authorize a device to communicate in the network. A successful activation will result in a valid session. Communication occurs while the session remains valid.

2.2.6.1 Activation Stage

The goal of activation stage is to have End-Nodes and Servers (both Network and Application) agree on Session Keys. This goal can be achieved through two different methods: Activation by Personalization (ABP) or Over the Air Activation (OTAA).

Barillaro, Sebastian; Rhee, Sokwoo; Kacker, Raghu; Badger, Mark; Kuhn, David; Escudero, Gustavo.
"Low-Power Wide Area Networks (LPWAN) for Communications of Mobile Sensor Data."
Paper presented at 2nd ACM/EIGSCC Symposium On Smart Cities and Communities (SCC 2019), Portland, OR, United States. September 10, 2019 - September 12, 2019.

In OTAA method, the End-Node and Network Server exchange a series of random numbers signed and encrypted with pre-shared keys, i.e., Network Key (NetKey) and Application Key (AppKey). After a successful handshake, both sides derive identical Session keys that will be used in Communication Stage. Session keys derived are known as Network Session Key (NetSKey) and Application Session Key (AppSKey).
In the ABP method, derived Session Keys (NetSKey and AppSKey) are pre-implanted in both sides before deploying End-Nodes. There is no handshake to derive Sessions Keys. When using the ABP method, it is unlikely that Session keys will ever change, making the sessions valid for unlimited time. Although implementation is easier, ABP method is considered less secure.

2.2.6.2 Communication Stage

Once End-Nodes are activated (with ABP or OTAA), they are authorized to communicate with Servers. With few exceptions, communications are initiated by End-Nodes, determined by the class of device. For Class A devices, an End-Node starts communication by transmitting a LoRa Packet. After a successful transmission, the End-Node opens a short receiving window to allow the Network to send a response. If no message is received, a second receiving window, which is longer, is opened. After that, it is up to the End-Node to decide when to re-start Communication Stage. It is assumed that End-Nodes are in sleep mode most of the time, with no way to receive any message from the Network during the sleep. If Network wants to communicate with End-Nodes, it has to wait until End-Nodes decide to start communication again, which may result in undetermined latency for down-link messages. Class B devices vary from Class A devices in that an End-Node opens up receiving windows periodically. Since it is assumed that End-Nodes has no accurate Real Time Clock (RTC) and they sleep during no-transmission, their receiving window timing should be synchronized by beacons sent by Gateways. Class B devices keep a balance between downlink latency and End-node energy consumption. The Class B category of devices was announced at the beginning of LoRaWAN Specification but implemented in version 1.1. Finally, Class C implements an always-open receiving window on End-Nodes. In this manner, latency can be reduced at the expense of End-Node's power consumption.

3 LPWAN Experiments at NIST

The LPWAN technology is designed for extreme efficiency in both distance and power consumption over the data rate. However, the data rate can be improved if the energy source is not limited. An experiment was designed to challenge the coverage and the data rate without restriction of energy source.
NIST has a shuttle service that transports people between the Shady Grove Metro Station and the NIST campus. The Shady Grove Metro Station is located 4.89 km (3.04 mi) *as the crows flies* from the NIST campus, as shown in Figure 2. The shuttle bus travels 10 km (6.2 mi) in 15 minutes each way. Most of the path is on local highways at speeds up to 90 km/h (60 mi/h). As shown in Figure 3 and Figure 4, this shuttle bus was used as a mobile object for the experiment and was equipped with a GSNN sensor. A Gateway and its antenna were installed on the roof of the tallest building (11 stories) on the NIST campus and in the vicinity area. Then, Network and Application server were installed on a laptop. Another computer with a large screen were used to show the location of the shuttle bus.

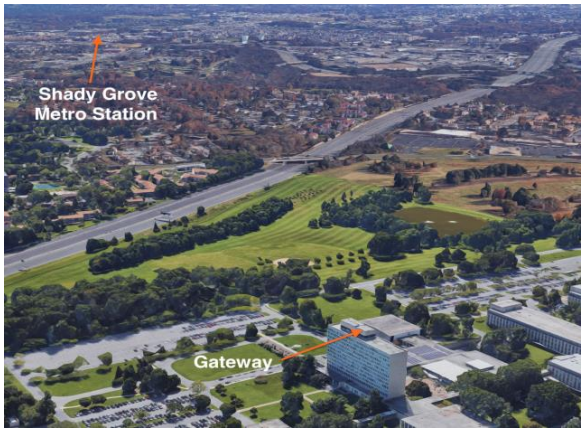


Figure 2: Simulated aerial view of the NIST Shuttle route area. The high-rise building in the foreground is where the gateway was deployed. In the background, the landmark points where the Shady Grove metro station is.

3.1 NIST LPWAN Facility Infrastructure

NIST LPWAN Facility was deployed using a LoRa/LoRaWAN Infrastructure. This deployment had been made before of LoRaWAN v1.1 specification was released. Thus, no Join-Server was deployed. A detail of infrastructure is described in Figure 3 and following sub-sections.

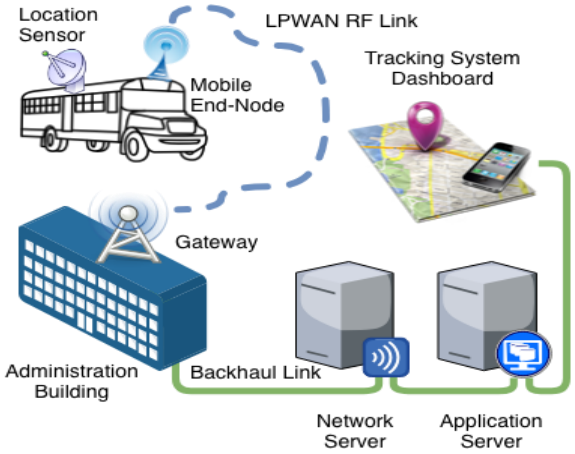


Figure 3: NIST LPWAN Testing Facility Infrastructure

Barillaro, Sebastian; Rhee, Sokwoo; Kacker, Raghu; Badger, Mark; Kuhn, David; Escudero, Gustavo.
"Low-Power Wide Area Networks (LPWAN) for Communications of Mobile Sensor Data."
Paper presented at 2nd ACM/EIGSCC Symposium On Smart Cities and Communities (SCC 2019), Portland, OR, United States. September 10, 2019 - September 12, 2019.

3.1.1 End-Node

Each End-Node was assembled by combining a development module, Pycom FiPy, with an expansion board including a GNSS sensor Pycom PyTrack (see Figure 4). An End-Node was connected to the vehicle power source of the shuttle bus, which eliminated the issue of limitation of power supply. The antenna attached to the End-Node was a 4 dBi water-proof omni-directional antenna with magnetic base and 1.5 m cable passing through the driver's window.

End-Node software was coded in micro-Python. End-Nodes transmit in any of the 64 channels allowed in ISM US-915 specification by default. The End-Nodes were configured to use only the transmission channels covered by The Gateway. Application ID (APPEUI) and Application Key (AppKey) were also set in the End-Node software configuration. After configuration, each End-Node executed a join handshake via Over-The-Air-Authentication (OTAA) until it received a join confirmation. Finally, it ran an infinite-loop where it calculated the GPS coordinates using the GNSS location sensor and transmitted it through LoRa/LoRaWAN.

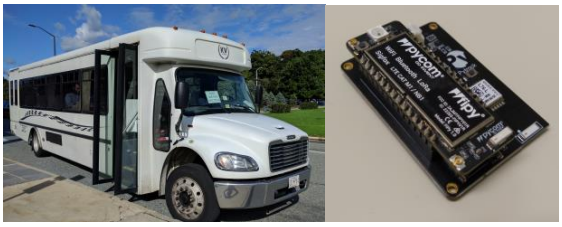


Figure 4: a) NIST Shuttle. b) Pycom FiPy + PyTrack Location Sensor

3.1.2 Gateway

The Gateway was built with a LoRa Gateway Module RisingHF RHF0M301 on top of a Raspberry Pi, as shown in Figure 5.

The LoRa Gateway Module was equipped with SX1301 + 2x SX1257, capable of covering one sub-band (8 + 1 channel) of 8 sub-bands (64 + 8) channels available in the ISM US-915 specification. The same type of antenna was used by both End-Node and Gateway. Packet-forwarder was provided by Gateway manufacturer, based on an implementation by Semtech. Packet-Forwarder was configured to communicate with Network Server. The Packet-forwarder was configured to use the operating frequency (US-915) and sub-band 1: channels 8 to 15 in 903.9 - 905.3 MHz frequency. Raspberry Pi OS (Raspbian) configuration was modified to meet the NIST IT Security policies. After that, the device was approved by NIST IT Security Officer and connected to the NIST-Net LAN Network.

This Gateway was installed inside a weather-proof box on the roof of the Administration Building which is eleven stories tall.

3.1.3 Network Server

The selection process of a Network Server was based on the same criteria as the choice of the LPWAN technology, i.e., no proprietary solutions, built-in security, possibility of ownership of the entire platform, and execution without Internet connection.

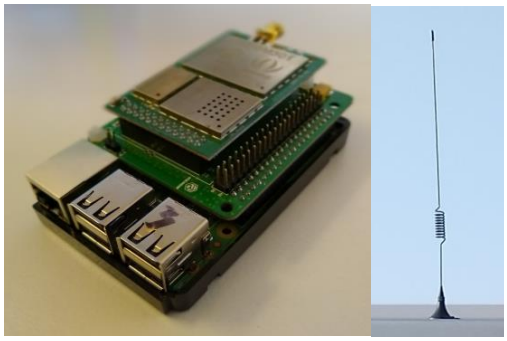


Figure 5: a) Gateway Module. b) Antenna (not in scale)

LoRa Server was selected as Network Server to be consistent with the criteria. LoRa Server software was deployed in a virtualized instance of Linux running on a Macbook Pro laptop.

Some confusion may arise from the nomenclature of LoRa Server components. Two main components in the LoRa Server project named Network Server and Application Server make up what LoRaWAN specification calls Network Server. In order to avoid ambiguity on nomenclature, "LoRaServer-" prefix is added when referring to LoRa Server project components. The functional responsibility of the network-server component (LoRaServer-network-server) is to de-duplicate and process uplink frames received by the gateway(s), to handle the LoRaWAN mac-layer, and to schedule downlink data transmissions [8]. LoRaServer-Application-Server is responsible for the device "inventory" part of a LoRaWAN infrastructure, handling join-requests as well as handling and encryption of application payloads [9].

All the administration and configuration are done via a web interface, as shown in Figure 6. Basic configuration includes: Registration and parametrization of Gateway(s), End-Node(s), and their integration to handle communication with LoRaWAN-Application-Server (the Application Server, according to the LoRaWAN specification nomenclature).

3.1.4 Application Server

The selection process of an Application Server software was based on the same criteria as the Network Server. InfluxDB [10] for data storage, combined with Grafana [11] for data visualization, were deployed in another virtualized instance of Linux, in the same Mac Laptop Host. Most of the InfluxDB configuration was done via the command-line interface except all the Grafana Configuration which was done using a web interface. A Grafana Worldmap Panel [12] was installed to show End-Point location in the map.

3.1.5 Tracking System Dashboard

A computer with a web-browser displayed on a large screen was demonstrated at SIM Week in September 2018 [13].

Barillaro, Sebastian; Rhee, Sokwoo; Kacker, Raghu; Badger, Mark; Kuhn, David; Escudero, Gustavo.
"Low-Power Wide Area Networks (LPWAN) for Communications of Mobile Sensor Data."
Paper presented at 2nd ACM/EIGSCC Symposium On Smart Cities and Communities (SCC 2019), Portland, OR, United States. September 10, 2019 - September 12, 2019.

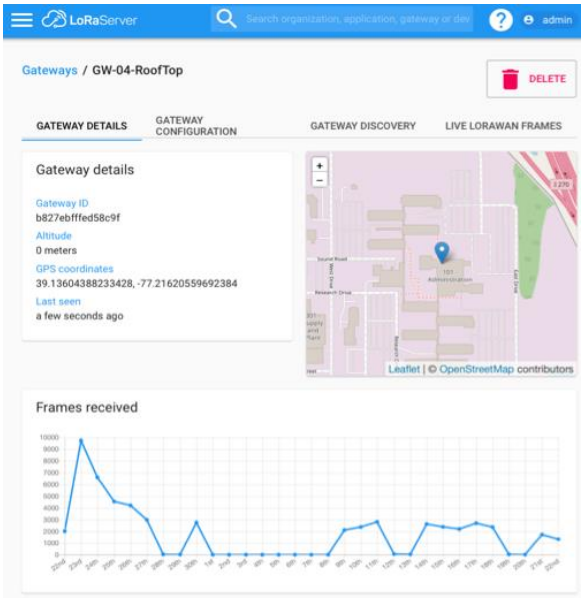


Figure 6: Gateway Status in LoRaServer. Altitude shows a wrong value due lack of GNSS sensor. Bottom graph shows frames received per day.

3.2 Experimental Tests

All tests were exploratory. The intention was to learn about LPWAN technologies from first-hand experiences, over and beyond technical references and marketing brochures. All the test should be considered preliminary and their results cannot be considered as conclusive.

3.2.1 Transmission frequency test (pre-deployment test)

This test consisted of transmitting empty packets (no payload) from the End-Node without requesting acknowledgement or delay between cycles in the loop. It was possible to receive a message almost every second under nearly ideal condition. For this test, the Gateway and the End-Node were placed inside the same room with metal walls.

3.2.2 NIST Shuttle route coverage test

Second test showed that the NIST shuttle bus could be tracked throughout almost the entire route. There were some blind spots where the transmissions were lost, possibly due to ground conditions, bridges, electrical transformers, etc. Figure 7 shows all locations where signal was received during a work day. Unlike the first test, the frequency of reception of messages decreased – once every 2 seconds to 4 seconds, if not more.

3.2.3 Latency Test

Another interesting characteristic to consider was the responsiveness of the system (Latency). Packets were received from a number of locations, but was the NIST shuttle actually at the location at the time it was displayed on the map? To answer this question, the location of the shuttle on the map was checked when the shuttle was approaching the Administration Building, while both the NIST shuttle and the map could be observed at the same time. It was confirmed that the location of the shuttle on the map was correctly displayed during this test. This was a special situation where the shuttle was moving very slowly. But it was impossible to answer that question while the NIST shuttle was traveling on highway and was out of sight. Although there are many steps involved between acquisition of the geographic coordinates by the GNSS sensors and display of the orange dot on the map, the current test solely focused on the end-to-end latency of the system as a whole.

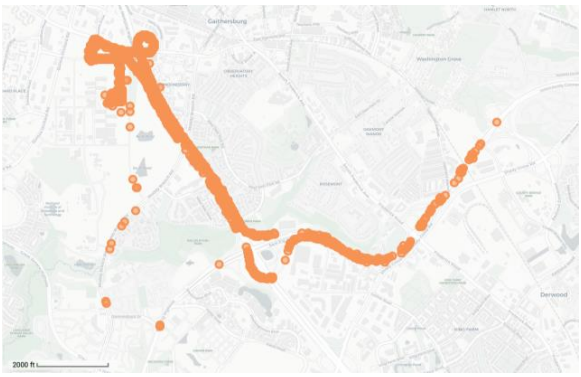


Figure 7: Accumulated locations received from the NIST Shuttle, including from unanticipated alternative routes.

In an LoRa/LoRaWAN infrastructure, the timestamp is assigned to a message by packet-forwarder software running in the gateway that receives the message. The gateway clock is synchronized with GNSS time, or with the Network Time Protocol (NTP) server in case that there is no GNSS Sensor. LoRa/LoRaWAN does not require end-nodes to have an RTC. This way, the end-nodes can be more inexpensive and more power efficient. If needed, an end-node may have an RTC to add a timestamp to sensor data. But this timestamp should be transmitted as part of the payload, which makes messages longer.

Our end-node was equipped with GNSS which could be used to add a timestamp to the location data. Unfortunately, Location Library offered by Location Sensor Manufacturer did not offer GNSS Time [14]. New libraries were developed by Pycom's user community to solve this problem, but they were not ready at the time of this test. Another option was to add a hardware RTC, but that would make the end-node more complex.

The problem was easily solved by using a separate reference system. A passenger equipped with a smartphone aboard the NIST Shuttle volunteered to help the study and reported independent real-time location data using a smart-phone LTE connection. The End-Node transmitted its location through the LPWAN setup in the NIST Facility as designed. Both location data sets were observed at the same time on the same screen. With the exception

Barillaro, Sebastian; Rhee, Sokwoo; Kacker, Raghu; Badger, Mark; Kuhn, David; Escudero, Gustavo.
"Low-Power Wide Area Networks (LPWAN) for Communications of Mobile Sensor Data."
Paper presented at 2nd ACM/EIGSCC Symposium On Smart Cities and Communities (SCC 2019), Portland, OR, United States. September 10, 2019 - September 12, 2019.

of some lost messages, it was possible to observe a good correlation between the two data sets, as shown in Figure 8.

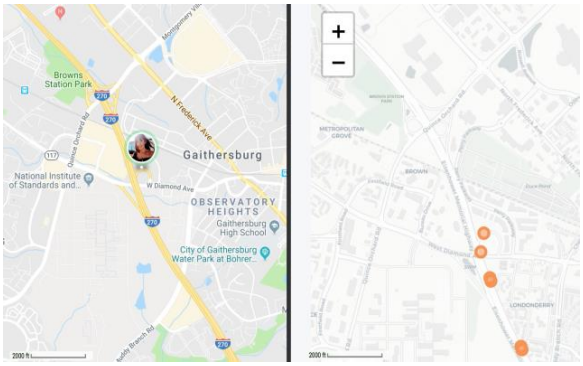


Figure 8: Side-by-side comparison. On the left, real-time location received through LTE connection. On the right, location received through the NIST LPWAN infrastructure. The upper orange dot shows the last received location, coinciding with same location reported through LTE link at that time.

3.2.4 Range test

Typical marketing brochures and white papers about wireless communications specify communication range in terms of physical distance. However, comparison of different technologies should use link-budget as the main parameter. Even with the same link-budget, communication range is seriously affected by terrain shape, weather condition, and existence of physical objects like buildings and trees that affect propagation of signal. A range test would tell us more about the context of the deployment than the technology itself.

The last test was conducted to check the communication range of a device in the LPWAN infrastructure at NIST which was deployed to cover the suburban area of Gaithersburg, MD. The End-Node was attached to a vehicle that drove around during a weekend in the vicinity of the NIST Campus. An analysis of the data during the test showed that a message from the farthest distance was received when the car was traveling on I-270 just before crossing Fall Rd Bridge, which was 8 km (5 mi) away from the gateway. Due to the characteristics of the landscape, most of the messages before that were lost even when they were transmitted at a shorter distance than the farthest location that a message was received from. The map in Figure 9 was configured to show all historical data, not just the last data point. As can be seen in Figure 9, an orange area covered almost the entire path of the Shuttle. The system ran continuously for a week.

4 Discussions

The paper described a tracking system using a LPWAN setup developed to analyze characteristics of the wireless technology. The experiment focused on a subset of important characteristics, i.e., transmission frequency, coverage, latency, and communication range. However, there are a number of additional characteristics that may have a significant impact on the overall performance of the IoT system based on the LPWAN technology. Examples includes, but not limited to, scalability, power consumption, robustness against interference, and co-existence with other types of wireless technologies. The timescale used in the latency test in this experiment was in the order of minutes and the measurement was done based on qualitative human observation. Future experiments may need to employ more quantitative measurement methods using more accurate references.

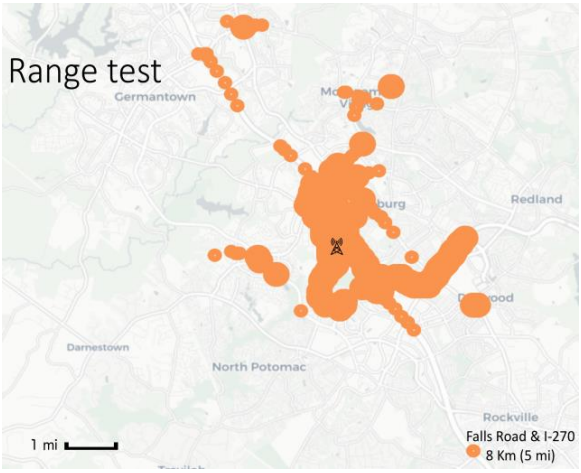


Figure 9: NIST LPWAN Infrastructure range test. The furthest received message was transmitted 8 Km far from the Gateway. Location of the Gateway is drawn on the map.

ACKNOWLEDGMENTS

Special Thanks to:

- Dhananjay *DJ* Anand, for invaluable help and support to set-up these experiments on time.
- Charles Prado, for his assistance to set up a demo.
- Jorge Carrasco, for his willingness to supervise the end-node while driving the NIST Shuttle.
- Maria Betania Antico, for sharing her smart-phone real-time location while traveling in NIST Shuttle as a volunteer passenger, and constant support in my professional development, and life.
- Nanita Yeboah, for carrying end-node during entire weekend to test LPWAN NIST Infrastructure range.
- Gurmindersingh *Gini* Khalsa, for his inspiring safety restrictions, source of inventiveness and creativity that guided me to find LPWAN.

Barillaro, Sebastian; Rhee, Sokwoo; Kacker, Raghu; Badger, Mark; Kuhn, David; Escudero, Gustavo.
"Low-Power Wide Area Networks (LPWAN) for Communications of Mobile Sensor Data."
Paper presented at 2nd ACM/EIGSCC Symposium On Smart Cities and Communities (SCC 2019), Portland, OR, United States. September 10, 2019 - September 12, 2019.

- Héctor Laiz and Osvaldo Jalon, for trusting my professionalism to represent INTI with world-class excellence abroad.

DISCLAIMER

Official contribution of the United States government; not subject to copyright in the United States. Certain commercial products may be identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by the National Institute of Standards and Technology, nor does it imply that such products are necessarily the best available for the purpose.

REFERENCES

[1] A. Ross, "Information Age's guide to the low-power wide area network (LPWAN) landscape," 27 June 2019. [Online]. Available: https://www.information-age.com/low-power-wide-area-network-lpwan-123483151/. [Accessed 25 July 2019].

[2] LoRa Alliance Technical Committee Regional Parameters Workgroup, "LoRaWAN 1.1 Regional Parameters," January 2018. [Online]. Available: https://lora-alliance.org/sites/default/files/2018-04/lorawantm_regional_parameters_v1.1rb_-_final.pdf. [Accessed 25 July 2019].

[3] LoRa Alliance Technical Committee, "LoRaWAN 1.1 Specification," October 2017. [Online]. Available: https://lora-alliance.org/sites/default/files/2018-04/lorawantm_specification_-v1.1.pdf. [Accessed 27 July 2019].

[4] LoRa Alliance Strategy Committee, "Geolocation Whitepaper," January 2018. [Online]. Available: https://docs.wixstatic.com/ugd/eccc1a_d43b3b29dfff4ec2b00f349ced4225c4.pdf. [Accessed 25 July 2019]

[5] B. Ray, "LoRa Localization," 30 6 2016. [Online]. Available: https://www.link-labs.com/blog/lora-localization. [Accessed 25 July 2019].

[6] I. Butun, N. Pereira and M. Gidlund, " Analysis of LoRaWAN v1.1 Security," in *MobiCom Mobile Computing and Networking*, Los Angeles, 2018.

[7] I. Butun, N. Pereira and M. Gidlund, "Security Risk Analysis of LoRaWAN and Future Directions," in *4th ACM MobiHoc Workshop on Experiences with the Design and Implementation of Smart Objects*, Los Angeles, California, 2018.

[8] "Loraserver documentation," [Online]. Available: https://www.loraserver.io/loraserver/overview/. [Accessed 25 July 2019].

[9] "Loraserver-Application-Server," [Online]. Available: https://www.loraserver.io/lora-app-server/overview/. [Accessed 25 July 2019].

[10] "InfluxDB," [Online]. Available: https://www.influxdata.com/time-series-platform/. [Accessed 25 July 2019].

[11] "Grafana," [Online]. Available: https://grafana.com. [Accessed 25 July 2019].

[12] "Grafana Worldmap Panel," [Online]. Available: https://grafana.com/grafana/plugins/grafana-worldmap-panel. [Accessed 25 July 2019].

[13] "SIM Week," September 2018. [Online]. Available: https://sim-metrologia.org/2018/09/21/sim-week-2018-gaithersburg-september-24-28-2018/. [Accessed 25 July 2019].

[14] Pycom, "PyTrack Documentation," [Online]. Available: https://docs.pycom.io/pytrackpysense/apireference/pytrack/. [Accessed 25 July 2019].

Barillaro, Sebastian; Rhee, Sokwoo; Kacker, Raghu; Badger, Mark; Kuhn, David; Escudero, Gustavo.
"Low-Power Wide Area Networks (LPWAN) for Communications of Mobile Sensor Data."
Paper presented at 2nd ACM/EIGSCC Symposium On Smart Cities and Communities (SCC 2019), Portland, OR, United States. September 10, 2019 - September 12, 2019.

Evaluating Sensor Algorithms to Prevent Kitchen Cooktop Ignition and Ignore Normal Cooking

Amy Mensch, Anthony Hamins, Z.Q. John Lu, Matthew Kupferschmid, Wai Cheong Tam, and Christina You
National Institute of Standards & Technology
Gaithersburg, MD 20899-8664

Background

Cooking equipment is involved in nearly half of home fires in the USA, with cooktop fires the leading cause of deaths and injuries in cooking-related fires [1]. While new electric-coil cooktops must pass the UL[1] 858 [2] "abnormal cooking test," which aims to prevent cooktop fires, there is no such requirement for older and other types of cooktops. In this study, we considered the use of gas and particle sensors to provide early warning and/or stop cooktop ignition of foods and oils. Thus, the objective of this study is to find new, data-driven ways to reduce the risk of cooktop fires. Our approach is to develop and test the performance of sensor-detection algorithms using threshold analysis and machine learning methods.

Experimental Methods

Measurements were made in a mock kitchen using both electric and gas cooktops. There were four different burners used in the experiments: the small 15 cm diameter electric coil heating element with a measured power of 1.1 kW, the large 20 cm diameter electric coil heating element with a measured power of 1.8 kW, the medium gas burner with an estimated heat output of 3.4 kW, and the large gas burner with an estimated heat output of 4 kW. Sensors were placed in the exhaust duct above the cooktop and exposed to the gases and particles representative of cooking.[2] The flow in the exhaust duct (15 cm diameter) was characterized using a velocity probe placed in the center of the duct about 20 diameters downstream of a bend. The typical average velocity was 3.4 m/s with a standard uncertainty of ± 0.1 m/s. The average velocity varied between experiments with a standard deviation of 0.2 m/s. Using the electric coil cooktop, the duct temperature increased by an average of 9 °C causing an estimated reduction in duct mass flow of 3 %. For the gas cooktop, the duct temperature increased by an average of 23 °C, which is estimated to reduce the duct mass flow by 7 %. Additional details of the experimental apparatus and methods are described in Ref. [3].

A previous series of electric-coil element cooktop experiments in the same mock-up kitchen monitored sensor performance during the heating of vegetable oils, water, hamburgers, and salmon [3]. Many of the oil experiments and one salmon experiment led to ignition. The data from those experiments were supplemented with data from additional ignition and normal cooking experiments covering a wide range of conditions. The additional experiments, described in Table 1, included cooking bacon, french fries, and chicken. For a few experiments, multiple pans were heated on separate burners simultaneously. Some experiments were conducted using a gaseous (methane) fueled cooktop. In total, 39 of the 60 experiments led to ignition of the oil or food.

[1] Underwriters Laboratories

[2] Future experiments will consider the effects of sensor placement, ventilation, and configuration.

Mensch, Amy; Hamins, Anthony; Lu, John; Tam, Wai Cheong.
"Evaluating Sensor Algorithms to Prevent Kitchen Cooktop Ignition and Ignore Normal Cooking."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

Table 1 Conditions for Experiments Augmenting Ref. [3]

Ignition	Heating Source	Pan Type	Pan Diameter	Food and Amount
N	electric coil	none	N/A	N/A
Y	electric coil	cast iron	20 cm	50 mL canola oil & 2 L water on separate burners
N	electric coil	cast iron & aluminum	20 cm	50 mL canola oil in each pan
N	electric coil	cast iron	20 cm	282 g chicken legs (2), 200 mL canola oil
Y	electric coil	cast iron	25 cm	223 g frozen french fries, 500 mL canola oil
N	electric coil	cast iron	25 cm	220 g bacon (8 slices)
Y	electric coil	cast iron	20 cm	110 g bacon (4 slices)
Y	electric coil	cast iron	20 cm & 25 cm	50 mL & 100 mL canola oil in separate pans
Y	electric coil	cast iron	20 cm & 25 cm	50 mL & 100 mL olive oil in separate pans
N	methane	none	N/A	N/A
Y	methane	cast iron	25 cm	100 mL canola oil
N	methane	cast iron	25 cm	N/A
Y	methane	cast iron	20 cm	50 mL canola oil
N	methane	cast iron	20 cm & 25 cm	50 mL & 100 mL canola oil in separate pans

Pan Temperature Measurements

In each experiment pan temperatures were measured at one or more locations using Type-K thermocouples either spot welded or peened to the top surface of the pan. The thermocouples showed significant variations in temperature across the pan surface. The standard uncertainty of the Type-K thermocouples was ± 3 °C. Figure 1 shows calibrated infrared (IR) images of dry (no oil) cast iron pans. The images reveal the distribution of temperature on the small electric coil element and on the large gas burner, which was influenced by pan orientation and geometry. The maximum temperature the camera could monitor was 370 °C, so regions above that temperature are shown as white. The simultaneous thermocouple measurements that were used to calibrate the IR images are labeled in the figure. From the thermocouple calibrations, the pan emissivity ranged from 0.88 to 0.96, depending on the experiment. The uncertainty in the IR temperatures was ± 8 °C. Figure 2 shows the pan surface thermocouple measurements during an experiment that led to ignition of canola oil. These figures demonstrate that temperature variation across the pan's bottom surface could reach 50 °C.

Figure 2 shows that the time series of the pan center temperature lagged temperatures measured toward the edge of the pan. Because the hottest region of the pan is where ignition is most likely, the maximum temperature was estimated in experiments when only the center pan temperature was measured. The estimate was based on a linear regression relationship between the thermocouple readings at the pan center and at the edge locations: 5 cm from the center in the 20 cm diameter cast iron pans, and 6 cm or 7.5 cm from the center in the 25 cm diameter cast iron pans. The linear regression relationships for the edge temperatures are shown in Table 2 as a function of center temperature for similar experiments (same pan size and burner size).

The average pan temperature at the time of ignition for all the experiments was 429 °C with a standard deviation of 25 °C. For the electric coil experiments, the maximum pan temperature at the time of ignition was between 403 °C and 483 °C. For the gas cooktop, the ignition temperatures of the pan were lower, between 371 °C and 382 °C. The gas cooktop also took much longer to

ignite. The average time to ignition was 536 s for the 25 cm pan on the large electric coil burner and 1104 s for the 25 cm pan on the large gas burner. Heating a pan represents a complex set of heat transfer processes involving radiation, convection and conduction, and the burner and pan configurations play an important role. Consistent with the slower temperature rise for the gas cooktop, ignition was not observed on the 20 cm cast iron pan with 50 mL of canola oil using the medium gas burner. For the 20 cm cast iron pan with 50 mL of oil, ignition occurred only when the pan was placed on the large gas burner.

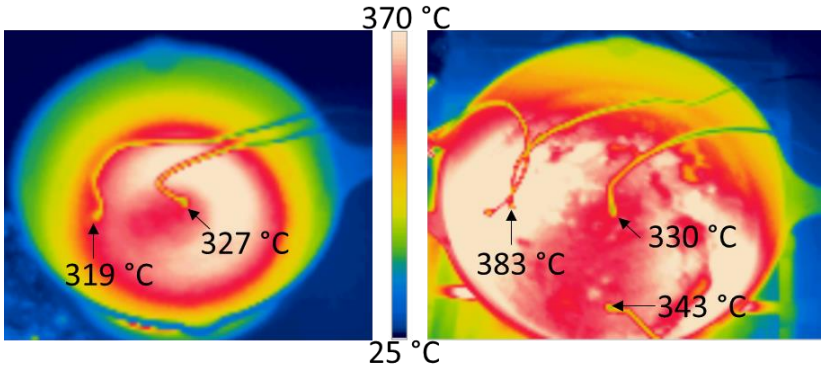


Figure 1. IR images showing the distribution of surface temperature of a 20 cm diameter cast iron pan heated by the small electric coil heating element (left) and the large gas burner (right).

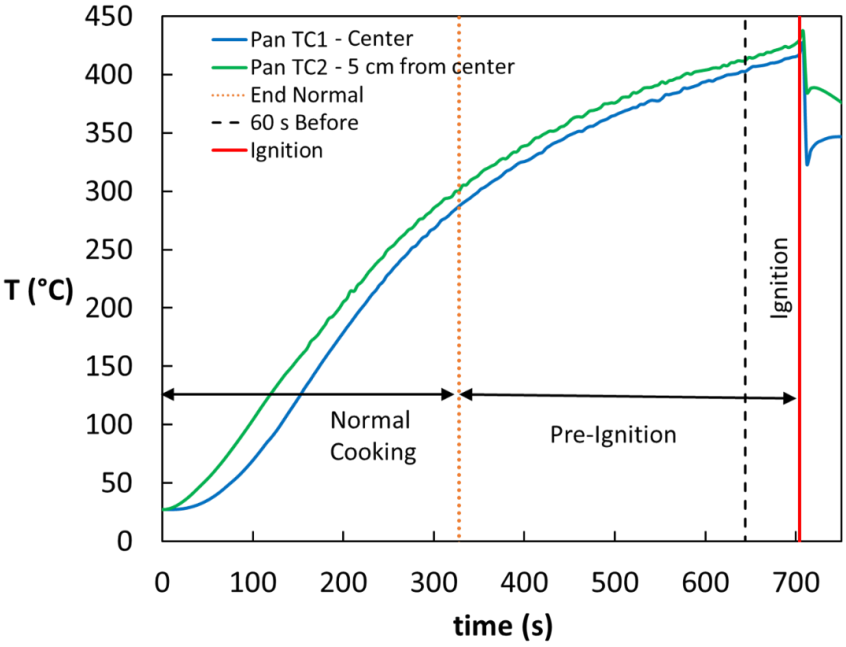


Figure 2. Pan surface temperatures and cooking regimes for an experiment leading to ignition of 50 mL of canola oil in a 20 cm diameter cast iron pan on the small electric coil heating element.

Mensch, Amy; Hamins, Anthony; Lu, John; Tam, Wai Cheong.
"Evaluating Sensor Algorithms to Prevent Kitchen Cooktop Ignition and Ignore Normal Cooking."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

Table 2 Relationships Between Pan Thermocouple Temperatures

Experiment Type	Linear Regression	R^2
Electric coil, cast iron 20 cm pan, small burner	$T_{5cm} = 0.972\ T_{center} + 23$ °C	0.99
Electric coil, cast iron 25 cm pan, small burner	$T_{6cm} = 1.07\ T_{center} + 16$ °C	1.00
Methane, cast iron 25 cm pan, large burner	$T_{7.5cm} = 0.967\ T_{center} + 31$ °C	0.99

Defining Normal Cooking

To develop an algorithm that predicts ignition, normal cooking must be defined. This is because sensor performance involves not only quantifying the rate of missed ignitions, but also the rate of false alarms. A missed ignition means a candidate algorithm would not have been triggered before ignition. To ensure there would be enough time for the algorithm to intervene and prevent ignition, the thermal lag of the cooktop and burner-pan system must be considered. Previous work suggests that a period of 60 s before ignition is enough time to intervene and prevent ignition [4]. A false alarm means that the algorithm predicts ignition is imminent, but the conditions are that of normal cooking, and ignition is not likely.

Figure 2 illustrates three periods of a typical experiment: normal cooking, pre-ignition, and ignition. Initially, all experiments started as normal cooking. At some point, the conditions exceeded some reasonable temperature-based or time-based limit and transition to "pre-ignition." Therefore, any condition that was not defined as normal cooking was labeled as pre-ignition, regardless of whether ignition eventually occurred later in the experiment. This was because the conditions during pre-ignition were considered beyond the requirements of normal cooking and potentially hazardous. Such conditions were accompanied by severely burned food and copious amounts of aerosol. The ignition period was defined as starting 60 s before ignition. Since there was no experiment in which the ignition period overlapped with normal cooking, it was possible for algorithms to predict ignition without interfering with normal cooking.

While the definition of the ignition period was straightforward, defining the reasonable limits of normal cooking required more nuance. The limits of normal cooking were based on either a maximum pan temperature, a safe food temperature, or the duration of cooking at an approximate pan temperature. For example, because the thickness of the vegetable oils and butter was thin (typically 3 mm), we assumed that the pan temperature gave a good indication of the oil temperature. When cooking foods such as meat, the pan temperature could be much hotter than the food, and food temperature was a better indicator of ignition potential than pan temperature. In defining normal cooking for meats, we took into account the USDA[3] safe minimum internal temperatures for chicken, 74 °C, fish, 63 °C, and ground beef, 71 °C [5].

The end of normal cooking for all types of oils and butter was defined when the pan temperature reached 300 °C. When deep-frying, it is recommended to keep oils below their smoke point, and the highest oil smoke points are around 230 °C [6]. Therefore, a limit of 300 °C allowed significantly more heating than recommended, while being well below oil ignition temperatures. For bacon, a USDA fact sheet states, "It's very difficult to determine the temperature of a thin piece of meat such as bacon, but if cooked crisp, it should have reached a safe temperature." [7]. Instead of relying on a "crispiness" determination, we treated bacon like oils, and the end of normal cooking was when the pan temperature reached 300 °C. This was reasonable since bacon is very high in fat, and liquid fat quickly coats the pan like vegetable oil. Photos taken at a pan temperature

[3] United States Department of Agriculture

of 300 °C showed that the bacon had already begun to blacken. Some bacon experiments led to ignition.

For chicken legs in 200 mL of preheated oil, the burner setting was on medium to maintain a pan temperature of about 200 °C for frying. The chicken legs were flipped every 4 min for a total cooking time of 18.5 min, which was 10 % longer than the time it took for the thermocouple inserted in the middle of the meat to reach 74 °C. This time was defined as the end of normal cooking, and the internal chicken temperature was 80 °C. For salmon fried in butter on high power for 4 min on each side, the thermocouples inside the meat did not show a steady increase in temperature. In most cases, the meat temperature exceeded 63 °C at least momentarily before the end of the 8 min of cooking, which was used as the end of normal cooking.

For hamburgers, the end of the frying procedure used by Cleary [8] was about 10 % longer than the time for the temperatures in the middle of the hamburgers to reach 71 °C. At the end of this procedure, the meat temperature was about 77 °C, which is an indication of well-done beef [9]. Therefore, the end of the frying hamburger procedure was defined as the end of normal cooking. For broiling hamburgers, the UL 217 Cooking Nuisance Smoke Test [10] specifies 25 min of broiling. However, in our experiments, adding an additional 10 % to the time when the hamburgers reached 71 °C, was less than 18 min (1122 s). This was defined as the end of normal cooking, and at this time the meat temperature was 82 °C.

For frozen fries in 500 mL of preheated oil, the burner power was adjusted periodically to maintain a pan temperature around 200 °C like was done for the experiments cooking chicken legs. There is no recommended safe temperature for fries, so the end of normal cooking was defined as 15 min of frying when the color of the fries had turned medium brown. After the end of normal cooking, the burner power was turned to high and the fries and oil later ignited.

Sensor Analysis

Sixteen sensors were positioned in the exhaust duct, approximately 3 m downstream of the range hood opening which was 0.8 m above the cooktop. The sensors monitored various quantities including CO_2, CO, temperature, humidity, smoke, hydrocarbons, alcohols, H_2, ammonia, natural gas, propane, volatile organic compounds (VOCs), and dust/aerosols. Raw data were acquired at 0.25 Hz. After each experiment, the average background signal for each sensor was subtracted from the raw signal output. Figure 3 plots the signals for a canola oil experiment on the gas burner, where the signals are normalized by the maximum value recorded from that sensor.

Sensor signal values and their ratios were evaluated to determine if a threshold value could be selected that both prevents ignition and ignores normal cooking conditions for all experiments. Machine learning was also used to develop algorithms to classify sensor data as representing normal cooking or pre-ignition conditions, and a similar performance metric was used. In addition to investigating the performance of thresholds of individual sensor values, we also considered the ratios between sensor values. CO_2 (PPM), duct temperature (K), and humidity (vol %) were used in the denominator of ratios. These signals did not include background subtraction to avoid dividing by zero because the values during the experiment were typically the same as the background.

Threshold Analysis

A threshold value of a sensor or sensor ratio could potentially miss ignitions as well as trigger false alarms. We considered the most conservative sensor or ratio threshold, which is the minimum

value obtained at least 60 s before all ignitions. The false alarm rate to evaluate the threshold performance was defined as the ratio of the number of experiments with a false alarm to the total number of normal cooking experiments. Table 3 summarizes the results of the threshold analysis, listing the best performing sensors and sensor ratios. The operating principle of the signal (in the numerator for ratios) is also listed. The sensor name reflects manufacturer product literature, but a sensor could respond to other things as well. For example, the dust optical sensor operates using light scattering, which can occur for both dust particles and cooking aerosols. Most of the sensors are not calibrated, and therefore, only the threshold voltage is reported. The indoor air quality sensor outputs in arbitrary units of PPM.

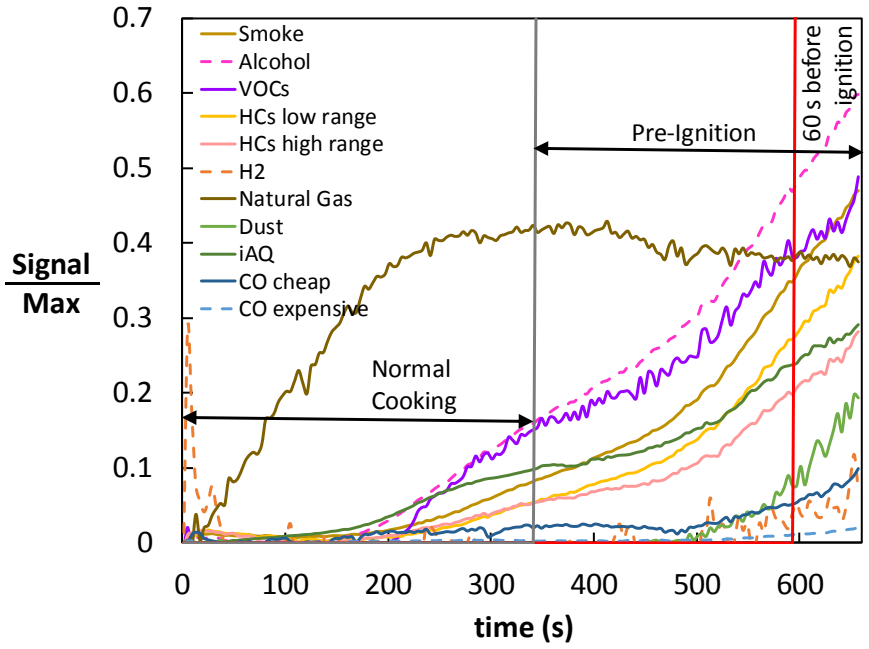


Figure 3. Sensor signals (background subtracted and normalized by sensor peak) and cooking regimes for an experiment leading to ignition of 50 mL of canola oil in a 20 cm cast iron pan on the large gas burner.

The best performance is for the volatile organic compounds (VOCs) sensor with a 2 % false alarm rate, or one false alarm in 60 experiments. The false alarm occurred about 2 min. before the end of normal cooking in one of the frying hamburger experiments with a 25 cm cast iron pan on the large electric coil burner. At that time, the thermocouples inside the hamburgers were both 68 °C, which is just below the safe temperature for ground beef (71 °C), but still within our definition of normal cooking. The ratios of sensors with duct temperature have similar performance to the sensor alone, but never better performance. One ratio that performs better than the sensor alone is the ratio of alcohol to humidity, which is slightly better than alcohol alone. A more significant improvement occurs for the ratio of the CO signals to CO_2, with the false alarm rate for the ratio of the expensive CO sensor nearly half of the CO sensor alone. The two different CO sensors both perform similarly despite the differences in output format (PPM vs. voltage) and price.

Mensch, Amy; Hamins, Anthony; Lu, John; Tam, Wai Cheong.
"Evaluating Sensor Algorithms to Prevent Kitchen Cooktop Ignition and Ignore Normal Cooking."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

Table 3 Threshold Performance of Sensors and Select Sensor Ratio Pairs

Sensor	Threshold	Units	False Alarm Rate	Operating Principle
VOCs	0.57	V	0.02	metal oxide sensor
VOCs / Duct Temp.	0.0019	V/K	0.02	metal oxide sensor
iAQ	12000	PPM	0.06	electrochemical
iAQ / Duct Temp.	40	PPM/K	0.06	electrochemical
VOCs / Humidity	0.28	V/vol %	0.07	metal oxide sensor
Dust	0.20	V	0.08	optical
Dust / Duct Temp.	6.3E-04	V/K	0.08	optical
iAQ / Humidity	6100	PPM/vol %	0.09	electrochemical
Alcohol / Humidity	0.65	V/vol %	0.11	electrochemical
CO, expensive / CO_2	0.012	PPM/PPM	0.11	electrochemical
Alcohol	0.92	V	0.12	electrochemical
Alcohol / Duct Temp.	0.0030	V/K	0.12	electrochemical
Dust / Humidity	0.081	V/vol %	0.13	optical
CO, cheap / CO_2	2.2E-05	V/PPM	0.16	electrochemical
Dust / CO_2	1.53E-04	V/PPM	0.16	optical
CO, expensive	4.3	PPM	0.20	electrochemical
CO, cheap	0.0078	V	0.22	electrochemical
CO, cheap / Duct Temp.	2.6E-05	V/K	0.22	electrochemical
CO, expensive / Duct Temp.	0.014	PPM/K	0.22	electrochemical
CO, cheap / Humidity	0.0060	V/vol %	0.23	electrochemical
CO, expensive / Humidity	3.4	PPM/vol %	0.24	electrochemical
Hydrocarbons, low range	0.22	V	0.25	electrochemical
Combustible gas & smoke	0.41	V	0.27	electrochemical
Combustible gas & smoke / Duct Temp.	0.0013	V/K	0.27	electrochemical
H_2	0.15	V	0.28	electrochemical

Machine Learning Analysis

The sensor signals were also used to train a multi-layer perceptron neural network to develop a model that differentiates between normal cooking and pre-ignition conditions. TensorFlow[4] was used as the application program interface (API) to implement machine learning. Two hidden layers with 64 neurons and 32 neurons were activated with a rectified linear unit (ReLU) activation function. A sigmoid activation function was used to calculate the output. Each time point was treated individually with a classification label, which was assigned 0 within the normal cooking window and 1 during the pre-ignition period. The method considered over 12 800 time points in the 60 experiments.

[4] Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the procedures adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Using a cross-validation method, the neural network model was trained using the data from 59 experiments and then tested on the last experiment. This process was repeated 60 times until each experiment was excluded from the training and used once as the test set. The output for the test experiment was a value between 0 and 1 for each time point, which was the model prediction for the probability of pre-ignition. The values were rounded to 0 or 1 and the predictions were compared to the labels from the experiment within the normal cooking and the ignition (60 s before ignition) windows. The overall performance was evaluated by counting the number of normal cooking data sets with any wrongly predicted values of 1 (false alarms) and the number of ignition data sets with any wrongly predicted values of 0 (missed ignitions). Predictions between normal cooking and the ignition window were ignored.

The missed ignition rate is the number of ignitions missed by the prediction divided by the number of experiments in which ignition was observed ($n_{ignitions} = 39$). The false alarm and missed ignition rates are given in Table 4 for a baseline case and cases with only single sensor input. The baseline case used 11 sensor signals as input to the neural network. These included the sensors listed in Table 3 as well as an electrochemical hydrocarbon sensor with a higher range and an electrochemical natural gas sensor. Table 4 summarizes the results of the neural network analysis. The sensor operating principles are also listed. The performance of the single sensor input cases are listed in order of performance (after the baseline case). The missed ignition rates are low because the neural network was trained to predict pre-ignition, which begins well before the ignition window. The best performing sensors in Table 4 are similar to the sensors with good threshold performance. The baseline model performs worse than many of the individual sensors, but better than the performance of the worst individual sensors (e.g. natural gas false alarm rate was 0.57 and missed ignition rate was 0.38). The baseline performance was probably negatively affected by including input data from the poorest performing sensors.

Table 4 Neural Network Model Performance of Baseline and Single Sensor Cases

Input Data	False Alarm Rate	Missed Ignition Rate	Operating Principle
Baseline, 11 sensors	0.50	0	
Indoor air quality (iAQ)	0.24	0	electrochemical
Dust	0.26	0	optical
Volatile organic compounds (VOCs)	0.26	0	metal oxide sensor
Hydrocarbons, low range	0.28	0	electrochemical
CO, expensive	0.33	0	electrochemical
Alcohol	0.35	0	electrochemical
CO, cheap	0.35	0.03	electrochemical
Combustible gas & smoke	0.39	0	electrochemical
Hydrocarbons, high range	0.43	0	electrochemical
H_2	0.43	0.03	electrochemical
Natural gas	0.57	0.39	electrochemical

Conclusions and Future Work

Threshold analysis and machine learning analysis were used to estimate the performance of individual sensors with a false alarm rate that was similar for both types of analysis. A precise and consistent definition of normal cooking versus pre-ignition was required to evaluate that

performance. To prevent ignition any algorithm must be triggered at least 60 s before ignition, which is defined as the ignition window. For the threshold analysis, the false alarm rate was reported for the threshold that was triggered before all ignition windows, so there were zero missed ignitions in every case. For the machine learning analysis, the rates of missed ignition were near zero since the neural network was attempting to detect pre-ignition, which began before the ignition window.

The performances of sensors with the lowest false alarm rates were in complete agreement between the two types of analysis. The best performing neural network models were based on sensors that also had good threshold performance. Although the machine learning false alarm rates were slightly higher than the threshold analysis, the initial neural network models were trained with only individual sensors as input. Future investigations of sensor performance will consider time-series effects to improve performance, such as evaluating sensor rate of change for threshold analysis and pre-processing input data to emphasize sensor rate of change for the machine learning analysis.

The combined information from multiple sensors was evaluated in a few limited cases: in sensor ratios with threshold analysis, and in the baseline neural network model with 11 sensor inputs. Some of the ratios performed as well as or better than the individual sensor values used in the ratios, but none of these cases performed better than the VOCs sensor alone. However, training neural networks with two or three sensors could provide an additional performance benefit by adding robustness and reliability to the model. Future work will involve combinations of two or three sensors as input data for neural network training, using the most promising sensors alone and in ratios: VOCs, iAQ, dust, CO, alcohol, duct temperature, humidity, and CO_2. The effects of transport conditions will also be considered.

[1] M. Ahrens, Home Fires Involving Cooking Equipment, National Fire Protection Association, Quincy, MA(2017).

[2] Underwriter's Laboratory, Northbrook, IL, Standard for Household Electric Ranges, Underwriter's Laboratory, Northbrook IL **UL 858** (2014).

[3] A. Mensch, A. Hamins, and K. Markell, Development of a Detection Algorithm for Kitchen Cooktop Ignition Prevention, *Suppression, Detection and Signaling Research and Applications Conference (SUPDET 2018)*, NFPA, Raleigh, NC(2018).

[4] E.L. Johnsson, Study of Technology for Detecting Pre-Ignition Conditions of Cooking-Related Fires Associated with Electric and Gas Ranges and Cooktops, Final Report, National Institute of Standards & Technology, Gaithersburg, MD(1998).

[5] Safe Minimum Internal Temperature Chart, United States Department of Agriculture, Food Safety and Inspection Service, https://www.fsis.usda.gov/safetempchart, (2019).

[6] Deep Fat Frying and Food Safety, United States Department of Agriculture, Food Safety and Inspection Service, https://www.fsis.usda.gov/wps/portal/fsis/topics/food-safety-education/get-answers/food-safety-fact-sheets/safe-food-handling/deep-fat-frying-and-food-safety/ct_index, (2013).

[7] Bacon and Food Safety, United States Department of Agriculture, Food Safety and Inspection Service, https://www.fsis.usda.gov/wps/portal/fsis/topics/food-safety-education/get-answers/food-safety-fact-sheets/meat-preparation/bacon-and-food-safety/ct_index, (2013).

[8] T.G. Cleary, A study on the performance of current smoke alarms to the new fire and nuisance tests prescribed in ANSI/UL 217-2015, National Institute of Standards and Technology, Gaithersburg, MD(2016).

[9] Meredith Home Group, Meat Temperatures Chart, https://www.marthastewart.com/270074/meat-temperatures-chart (2019).

[10] Underwriter's Laboratory, Northbrook, IL, Standard for Safety Smoke Alarms, **ANSI/UL 217** (2015).

Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard

Wai Cheong Tam[1], Thomas Cleary[1], Eugene Yujun Fu[2]
[1]*National Institute of Standards and Technology, Gaithersburg, MD, USA*
[2]*Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong*

Abstract

This paper presents a learning-by-synthesis approach to facilitate the utilization of a machine learning paradigm to enhance situational awareness for fire fighting in buildings. An automated Fire Data Generator (FD-Gen) is developed. The overview of FD-Gen and its capabilities are highlighted. Using CFAST as the simulation engine, a time series for building sensors including heat detectors, smoke detectors, and other targets at any arbitrary locations in multi-room compartments with different geometric configurations can be obtained. An example case is provided. Synthetic data generated for a wide range of fire scenarios from the example is utilized in a supervised machine learning technique. Preliminary results demonstrate that the proposed models can help to predict building fire hazards in real-time.

Keywords: machine learning, simulated data, fire fighting

Introduction

The rapidly expanding availability and diversity of Internet of Things (IoT) technologies revolutionize how observations and data measurements can be made in buildings with little to no human intervention. Interconnected terminal equipment and facilities, such as smart sensors, embedded actuators, mobile devices, and industrial systems, enable automated collection of specific data associated with the real-time conditions of both occupants and building spaces. Data can be systematically gathered and securely transmitted to desired destinations through various wireless or wired networks. The use of these data is believed to be a promising solution to overcome the current difficulties for fire fighting [1]. But, in spite of the recent efforts developed to the IoT technologies in providing both reliable data acquisition and efficient data transfer, none of the real-time data has been utilized in any significant degree by the fire safety community, particularly in the area of firefighting, where the lack of information on the fire-ground is known to be one of the leading factors to incorrect decision making for fire response strategies in which can exacerbate property losses and/or casualties.

Instantaneous data extraction to reveal trends, unseen patterns, hidden relationship, and/or new insights is computationally complex. In a typical building environment, a vast amount of data can be generated from various devices/systems. Since the data is collected from a wide range of sources, the characteristics of the data are highly unstructured and heterogenous [2]. Given a specific situation, there will be a need for faster response to the data. When a larger building environment is involved, the data volume, variety, and velocity (response requirement) increase at scale. Due to the nature of the data itself (high volume, high variety, and high velocity), traditional

Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

data analytic algorithms, such as predictive modeling and statistical analysis, are incapable of providing any constructive insights within seconds [3]. For that, a robust and computationally efficient algorithm that can analyze both structured and unstructured data to provide trustworthy insights into decision-making processes regardless of source, size, and type is vital for enhancement of situation awareness, operational effectiveness, and safety for fire fighting.

The machine learning paradigm (ML) [4] is among the top methods that can provide real-time prediction. Chenebert et al. [5] provided a decision tree (DT) classifier being trained based on image data for flame detection in outdoor environments. Using imagery as training data, Yin and his co-workers [6] developed a deep neural network (DNN) framework for smoke detection and they provided hand-crafted features that can help improve detection rate above 96% on their image dataset. Recently, more advanced ML architectures are proposed. A well-established fast object detection convolutional neural network (CNN) with 12 layers was used to identify flames for high-resolution videos given in residential building settings [7]. Aslan et al. [8] attempted to use the state-of-the-art ML architecture, generative adversarial networks (GANs), to classify flame and non-flame objects for videos from surveillance cameras. Also, a saliency detection method [9] based on DNN was proposed to provide wildfire detection for unmanned aerial vehicle and reliable accuracy was reported. Results from these research works show that the use of ML paradigm can overcome the real-time fire predictions that statistical based [10] and physics-based models [11] might not be able to handle. However, to best of our knowledge, the use of ML paradigm for data in time series in building fires does not currently exist.

Indeed, the primary challenge is the scarcity of real-world data for building sensors with fire events. The data problem has been raised in different literature [12]. For the fire safety community, it can be noted that acquiring the desired sensor data is not trivial because 1) fire event do not happen frequently, 2) time series data associated with fire event in building environments are not available to the public data warehouse [13], and 3) physically conducting full-scale fire experiments in buildings is costly and time-consuming. Moreover, no prior research work has been carried out to provide guidance on the data requirement for ML applications. With that, there can be a high possibility in which the experimental data is not usable. When a conventional ML paradigm demands a large amount of training data (in the order of million sets), we, propose a learning-by-synthesis approach to generate simulated data to facilitate the use of ML paradigms for prediction of building fire hazards. The research outcome is intended to enhance situational awareness for firefighting in buildings.

In the following sections, the overview of Fire Data Generator (FD-Gen) and its capabilities will be presented. An example case will be given. Preliminary results for the example will be shown and key finding will be highlighted.

Overview of FD-Gen and its capabilities

Fire Data Generator (FD-Gen) is a computational tool with its front-end written in Python[1]. The code is developed to generate a time series for typical devices/sensors (i.e. heat detector, smoke

[1] Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the procedures adequately. Such identification is not intended to imply recommendation or endorsement by the National

Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

detector, and other targets) in user-specified building environments. There are four fundamental elements associated with FD-Gen: 1) basic input handler, 2) distribution function module, 3) sampling module, and 4) simulation engine. It should be noted that the code is written to execute multiple runs with different configurations on the Fire Research Division computer cluster at NIST in parallel. Depending the availability on the computational resources, a maximum of ten thousand simulation cases can be completed in a single day.

Basic input handler

The growth and the spread of a fire depend on many factors. For FD-Gen, the basic input handler is capable of handling input parameters accounting for the effects associated with different geometric and/or thermal configurations for the building and various types of fires. Table 1 provides the overall statistics and the assumed distribution functions for a list of important parameters that were found in a previous study associated with residential buildings [14]. Users can utilize the information from the table as the range of inputs to configure the simulation cases. User-specified values will be needed in case of missing statistics or assumed distribution functions.

In addition to the parameters shown in the table, other input parameters include: a) the opening condition, wall density and wall specific heat of the buildings; b) the peak HRR, total energy and its plateau for fires; and c) the thermal properties, locations, and the normal direction for detectors. For commercial buildings, the current version of FD-Gen does not have the statistics and the assumed distribution functions for any input parameters. For that, user-specified values are required. However, this information can be obtained from [15] and might be considered in the future updates.

Parameter	Units	Min	Mean	Max	Assumed Distribution Function
Floor area	m^2	12.8	18.2	34.6	Lognormal
Ceiling height	m	2.13	2.61	3.66	Lognormal
Opening width	m	0.81	2.03	3.24	Lognormal
Opening height	m	1.93	2.27	NA	Lognormal
Wall conductivity	W/m-K	NA	1.03	NA	NA
Wall thickness	mm	13.5	14.3	15.9	Lognormal
Fire location[2]	--	1	2	4	Normal

Table 1: Statistics on selected input parameters for residential buildings in FD-Gen.

Distribution function module

In order to produce sampling values of the input parameters based on their assumed probability distribution, SciPy [16] is implemented. SciPy is a module written in Python that consists of a library with well-known probability distribution functions. In the current version, the code can account for the following univariate distributions such as Binomial, Exponential, Logistic, Lognormal, Normal, Poisson, Triangular, Uniform, Weibull, Gamma, and Beta. Based on recent literature [14], the Normal, Lognormal, and Uniform distribution functions will mostly be used.

Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

[2] The values indicate location of the fire with 1 = in the center of the room, 2 = against a wall, and 4 = at a corner.

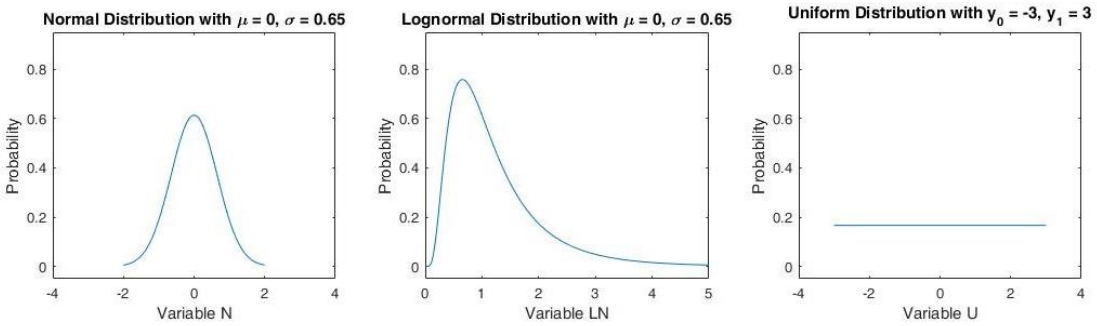
Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

The mathematical formulation of these distribution functions is provided below and Figures 1 show the profiles for the 3 distribution functions for reference:

$$Normal(x;\ \mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

$$Lognormal(x;\ \mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{[\ln(x)-\mu]^2}{2\sigma^2}} \tag{2}$$

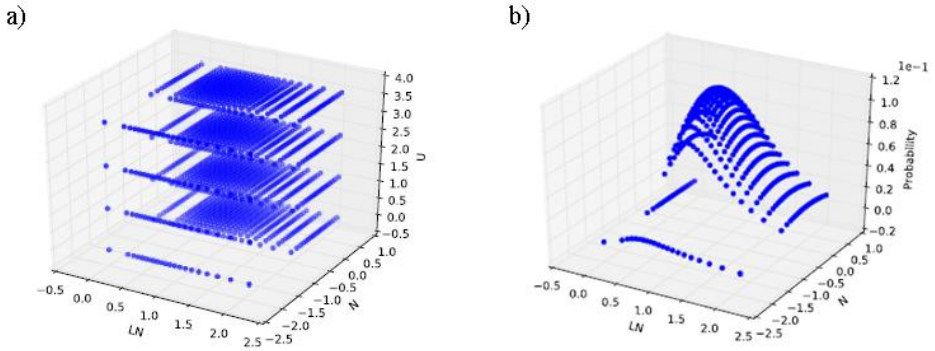
$$Uniform(x;\ y_0,y_1) = \frac{[G(x-y_0) - G(x-y_1)]}{y_1} \tag{3}$$



Figures 1: Profile for the 3 distribution functions with given μ, σ, y_0, and y_1.

Sampling module

In FD-Gen, a sampler connects a set of variables to their corresponding distributions and produces a sequence of points in the input space. This is an important feature for FD-Gen as it would greatly affect the cost of computational resources. For this current version, the code only supports 1 sampler and it is the grid sampler. Using Eqns. 1 to 3 with μ, σ, y_0 and y_1 being 0, 0.5, -3, and 3, respectively, Figures 2 show the sampling coverage in 3-D space. For the grid sampler with 20 points over the Normal and Lognormal distributions equally and 4 points over the uniform distribution equally spaced in the variable values, a total number of 1600 sampling points will be needed. In the future, Monte Carlo and/or Latin Hypercube Sampling can be implemented to provide the sampling flexibility and enhance numerical efficiency.

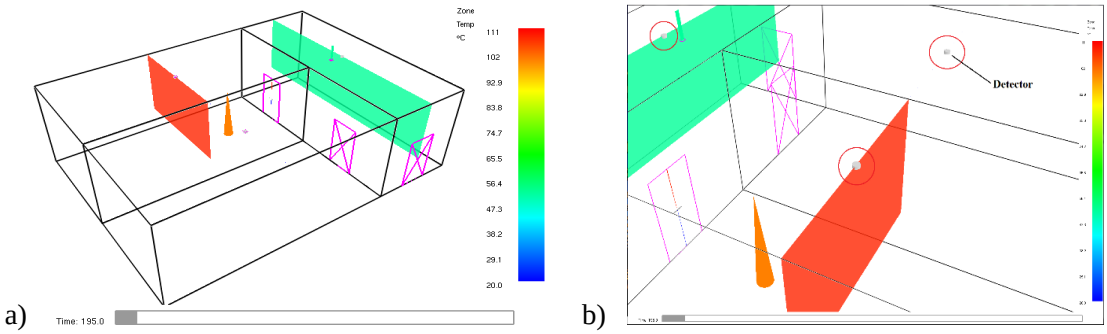


Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

Figures 2: a) 3-D dispersion of the sampling point coordinate and b) the probability map.

Simulation engine

CFAST [17] is used as the simulation engine for FD-Gen. In general, CFAST is a fire simulation program that divides compartments into two zones. Each zone includes a gas mixture/soot medium bounded by a ceiling or a floor, and four surfaces. Thermal conditions of each zone are assumed to be uniform. When there is a fire, a hot layer will form and the medium can be divided into an upper layer and a lower layer. If the fire persists, the upper layer increases in depth and the temperature will rise. When openings exist, there will be natural flow through the openings allowing air exchange between different compartments and zones. Figure 3a shows a typical simulation case with a user-specified fire in a zone for 3-compartments building structure.



Figures 3: a) Visualization of a typical CFAST simulation and b) exploded view for detectors.

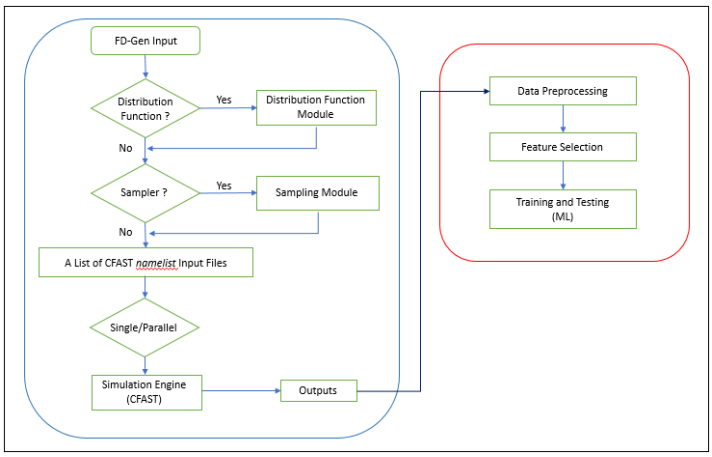


Figure 4: The workflow for FD-Gen.

In CFAST, devices, such as heat detectors, smoke detectors, and other targets can be specified. Although the CFAST simulation is based on the zone method, empirical equations are implemented to determine the local parameters such as gas temperature, smoke concentration, and total radiative heat transfer for heat detectors, smoke detectors, and other targets, respectively. Most importantly, CFAST is verified and validated such that the simulation results obtained from the program are reliable for a wide range of input conditions. Physically, verification ensures a

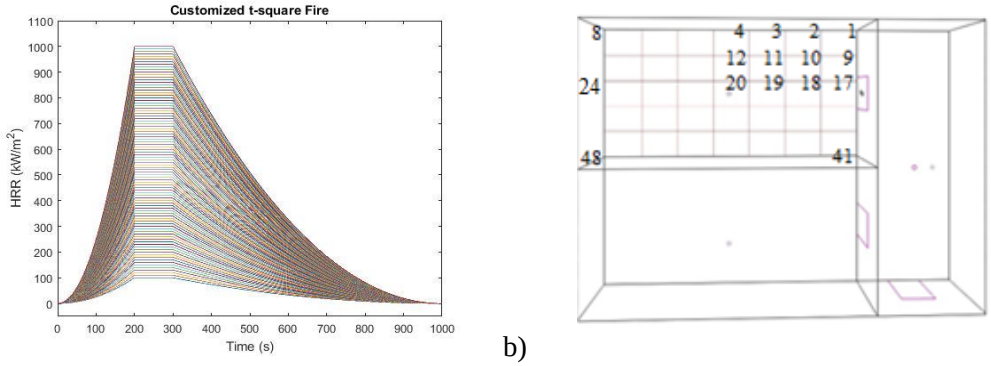
Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

given model is translated correctly into the computer program, and validation ensures the model accurately represents the phenomena of interest. Appendix A provides the full validated range of input parameters and Appendix B shows the modeling uncertainty for all output quantities of interest. Based on Appendix B, it is worth noting that CFAST is conservatively biased, in comparison to the experimental data, for all output quantities except those underlined fonts.

In summary, Figure 4 demonstrates an overall workflow for FD-Gen. An example case is provided in the next section for reference.

Preliminary Results and Discussion

Consider a single-story building with three compartments similar to that shown in Figure 3a. There are two offices and one corridor. The dimensions of the two offices are identical and they are 8m x 4m x 3.5m for length, width, height, respectively. The corridor dimensions are 3m x 8m x 3.5m. The material of the surfaces, ceiling, and floor is concrete. The office located in the upper left-hand corner is denoted as Office1 and the other office is denoted as Office2. It can be shown that there are 3 openings: one is in between Office1 and corridor, one is in between Office2 and corridor, and one is connected for corridor and the outdoor environment. The dimensions of the three openings are 2m in width and 1m in height. Initially, the openings are all closed, when the normal incident heat flux to the center of the opening is greater than 2.5kW/m^2, the opening will then be opened. Heat detectors are located at the center of the ceiling in each compartment. The outdoor conditions are normal with the temperature maintained at 20˚C at 100 kPa.



a) b)

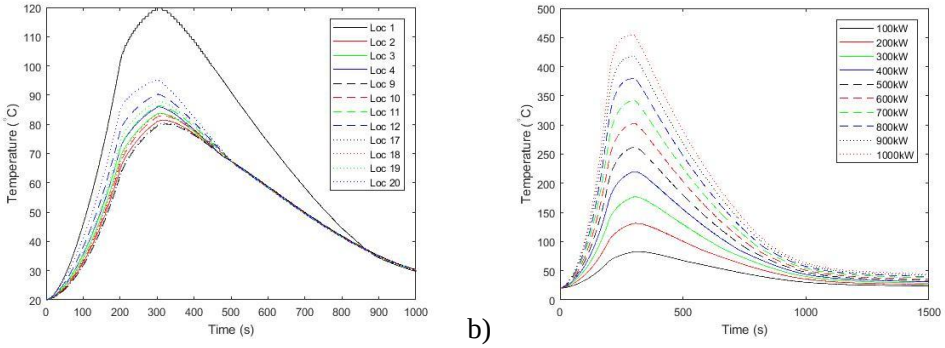
Figures 5: a) 100 t-square fire profiles and b) 48 fire locations in Office1.

Given the compartment settings provided above, simulation runs are executed for a fire with a wide range of peak heat release rates (HRRs) at different locations in Office1. Figure 5 shows the profile of 1000 HRR curves for fires and the 48 locations of a fire with a given HRR curve. Using a t-square relationship provided in CFAST and the limits of peak HRR (i.e. the lower bound to be 100 kW and the upper bound to be 1000kW), the varying HRR curves can be generated using grid sampling (with 10 kW increment) on peak HRR. Similarly, since the floor dimensions of Office1 are specified, the 48 locations of a fire with the HRR curve are generated using the grid sampler. It can be seen that 8 fire locations are specified in the horizontal direction and 6 fires are specified in the vertical direction. In total, detector temperature for 4800 simulation runs are obtained. The

Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

simulation time is set for 3600 s and output intervals are set to be 1 s. For this example, the total computational time is approximately 1.5 h.

Figures 6 show the detector temperature profiles for a 100 kW peak HRR fire for all 48 locations in Office1 and the detector temperature profiles for all t-square files at a location. The relative distance in between the fire location and detector location is 1.71 m in horizontal direction and 1.20 m in vertical direction (Loc 10). In Figure 6a, that peak detected temperature for Loc 1 is the highest. The observation is probably due to the corner effect associated with the fire. It should be noted that other output quantities such as smoke concentration and total incident heat flux can be obtained. For simplicity, this paper will only focus on temperature at selected locations.



a) b)

Figures 6: Temperature profiles a) for a fire with 100kW peak HRR at different locations and b) for fires with different HRR ranging from 100kW to 1000kW at Loc 10.

Machine Learning

Given a synthetic dataset, the use of the machine learning paradigm requires three additional processes: 1) preprocessing, 2) feature extraction, and 3) training and evaluation. Since the focus of this paper is on FD-Gen, only a brief discussion on the use of the machine learning paradigm will be included. It should be noted that the information provided in the subsections is only for demonstration. Actual development of a machine learning model will require more careful consideration. For simplicity, the example below would only take into account the data generated for cases with the same HRR profile (100 kW in Figure 6b) for 48 different fire locations (refer to Figure 5b). In total, the dataset contains detector temperature profiles for 48 simulation runs.

Labels	Location Range (m)	Number of profiles
Class 1: very close to the detector	0 - 1	2
Class 2: close to the detector	1 - 2	4
Class 3: a distance from the detector	2 - 3	6
Class 4: far away from the detector	3 - 4	6
Class 5: very far away from the detector	4 - 4.75	4
Class 6: at corner	4.75	2

Table 2: New labels for the example dataset.

Preprocessing

Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

There are three important matters that the data preprocessing will have to consider in this example case. The first one is to categorize target labels such that a machine learning model can be trained efficiently. Figure 6a shows that the change of detector temperature profiles depends primarily on the relative distance between the fire and the detector. The figure shows the closer the fire to the detector, the higher the maximum detector temperature. An exception is observed for the case with a corner fire. Physically, this is due to the corner effect of the fire that yields a ceiling jet with much higher temperatures. Based on the data behavior obtained from the data analysis and numerical experiment, the target labels for the current dataset will be modified and categorized into different "classes". Table 2 shows the breakdown information about the new labels. For labels in *Class 1*, the fire is located in a distance ranging from 0 m to 1 m.

The second matter is the removal of duplicate information in the dataset. As shown in Figure 5b, symmetry along the centerline across the horizontal direction of Office1 is observed. Since CFAST is a zone model with the identical boundary and initial conditions for the cases, the resulting detector temperature profiles for two separate fire cases (even with same HRR profile) at two different locations (i.e. Loc 12 and Loc 28) will be identical. This type of duplicate information is problematic when a machine learning model is being evaluated for prediction accuracy. Physically, this can be explained with the fact that a well-trained model based on a dataset will have perfect predictions (~100 % accuracy) when applied to a dataset used for training. However, when the model is applied, the prediction accuracy for such model might be very low. This behavior is considered as *overfitting*. To counter overfitting, unseen data will have to be used for testing [18]. For that, only data associated with fire locations from Loc 1 to Loc 24 will be considered.

Lastly, a segmentation on the dataset is required for real-time applications. Specifically, the entire detector temperature profile for a specific case as shown in Figure 6a will have to be divided into smaller time series. It can be understood that a machine learning model will be able to encode more precise information (i.e. trend, rate or change, etc.) about the fire using the entire detector temperature profile as an instance for a feature vector. In this scenario, it is highly likely that the model will have better prediction performance. However, such model would lose the ability of real-time detection, as it needs to wait for the complete temperature profile. In this example, 3600 s of data are needed. In order to build a model that is feasible to be used in real-time, the dataset is suggested to be divided into smaller segments using a rolling time window. In this scenario, only partial information about the temperature profile will be needed. In this pilot study, a parametric study will be carried out to examine the prediction performance for different lengths of the time window, *W*. The window size will vary from 0 s to 120 s (with increment of 10 s).

Feature Extraction

Feature extraction is a critical process for obtaining feature vectors that can include important information about the data. A good feature extraction will also help increase the prediction performance for a machine learning model. In this example, we aim to extract features based on the temperature profiles for fire location detection. For each fire event, we describe the temperature signal as: $T = [T_1, T_2, \dots, T_n]$, where T_i is the detected temperature at time stamp, *t*, of i. In addition, we further extract speed[3] signal: $S = [T_1/t_1, T_2/t_2, \dots, T_n/t_n]$, which indicates the average temperature growth rate ($^{\circ}\mathrm{C}/s$). Statistical features such as maximum and mean of these signals

[3] It is defined similarly as the average rate of change for a total time, *t*.

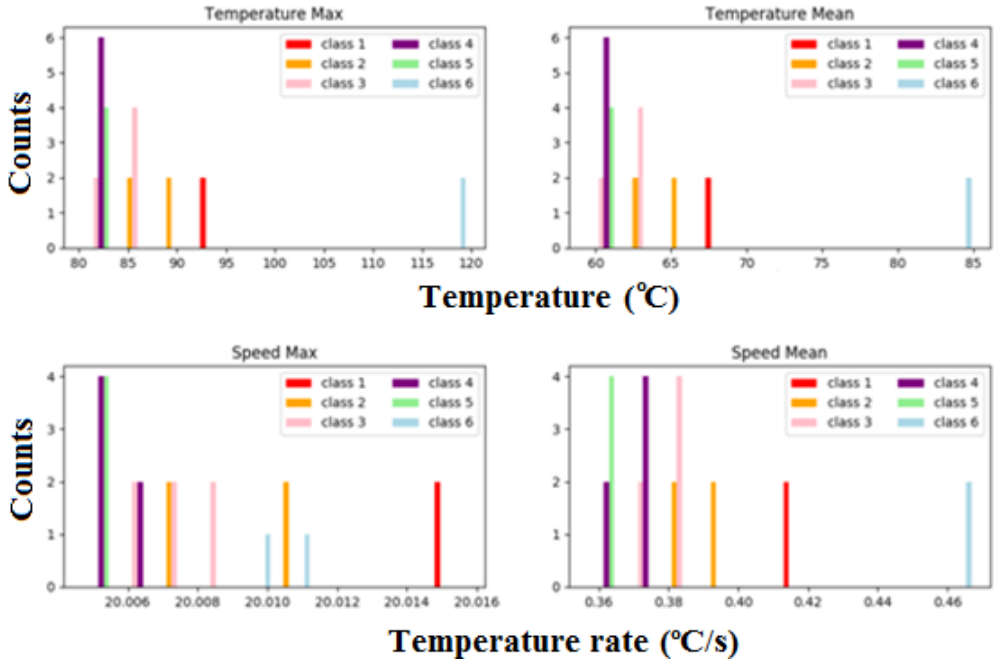
Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

are also extracted. Moreover, since the data is given in time series, the first order derivatives of these signals are also considered. Table 3 summarizes the signals and features that will be used in our model.

Signal	Features
Temperature	Maximum, mean, median
Speed	
First order derivative of Temperature	
First order derivative of Speed	

Table 3: Signals and their features.

Figure 7 shows the data distribution for the statistical (max and mean) temperature and speed signals based on the entire time series and four distinct behaviors are observed. It can be seen that 1) the fire located at the corner (*Class 6*) has a much larger maximum (refer to upper left plot) and average temperature (refer to upper right plot). The fire that is very close to the detector (*Class 1*) has the largest maximum temperature growth rate (refer to lower left plot). And for the fire that is far (*Class 4*) and very far away (*Class 5*) from the detector, they both have a lower maximum and lower average temperature (refer to upper plots). Cases associated with fire of *Class 5*, however, tend to have lower average temperature growth rates than that of *Class 4*. Figure 8 shows the data distributions for the statistical signals for temperature and speed segmented based on the rolling window technique. In this figure, the window size is 120 s. Similar to that of seen in Figure 7, clear distinctions associated with each feature are observed. For this purpose, it can be expected that the use of these features for building a model will likely lead to good performance.

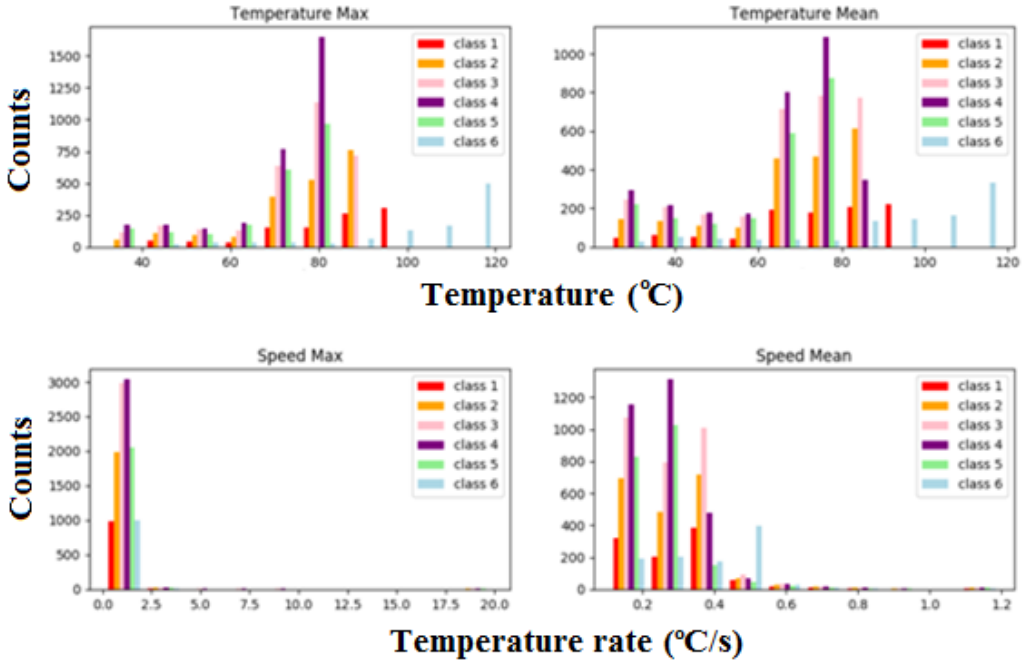


Figures 7: Data distribution for scenarios using the entire profile.

Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

Training and Evaluation

Given a dataset with feature vectors (refer to Table 3) and the corresponding labels (refer to Table 2), we can use machine learning paradigm to classify fire relative location. In this example, we build two machine learning models, one for Random Forest and one for Decision Tree. These models are used because of numerical efficiency and capability to handle the complexity of the problem. For problems with more complex data (i.e. prediction of tenability for another room at a given time), more advanced models, such as Support Vector Machines and/or neural networks, can be used.



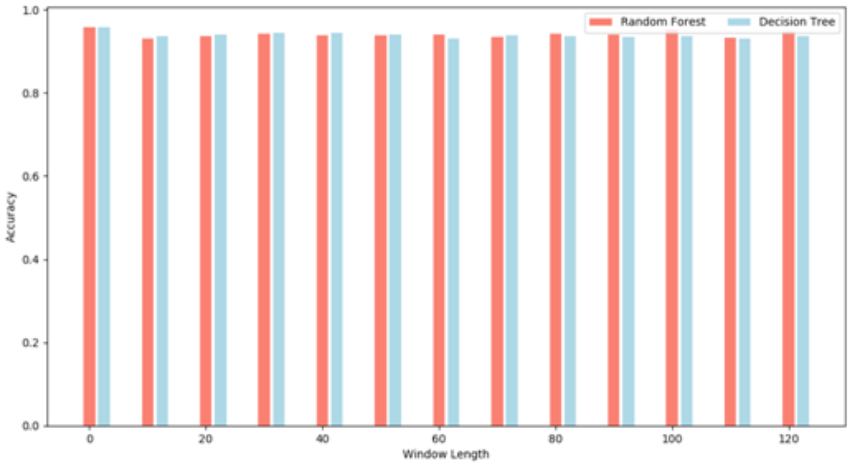
Figures 8: Data distributions for scenarios with real-time detection (W = 120 s).

For evaluation purposes, we divide the dataset into two subsets: a training set and a testing set. We train the detection model based on the training set and evaluate the performance of the model using the testing set. Since the dataset for this example is relatively small, we adopt leave-one-experiment-out cross-validation to evaluate our model. Specifically, we divide the dataset into 24 subsets where each subset contains only the data instances from one particular simulation run. We then train the classifier with the 23 sets among them (the training set) and carry out evaluation on the remaining one (the testing set). This is repeated 24 times for all subsets. The overall average accuracy across the 24 evaluations is reported as the final performance.

Figure 9 shows the model performance for use of the entire temperature profiles (denoted as $W = 0$ s) and the use of rolling window technique with different window sizes (denoted as $W > 0$ s). Preliminary results indicate that our proposed machine learning models are capable of detecting fire relative location with reasonable accuracy. For the scenario using the entire profiles, the

Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

accuracy[4] is 95.83%. Only one instance is wrongly classified: a C*lass 2* (close to the detector) instance is wrongly detected as C*lass 3* (medium distance to the detector). For the scenario using the rolling window technique, our models achieve reasonable accuracy even with partial information. The best performance is observed to be 95.08% and that is for W = 100 s using Random Forest. In general, these preliminary results demonstrate promising results for using machine learning paradigm to detect fire location based on detector temperature profiles in real-time.



Figures 9: Performance of fire location detection for the scenario of using the entire profile ($W = 0s$) and the scenario of real time detection ($W > 0s$).

Conclusions

The workflow of the Fire Data Generator (FD-Gen) is presented. The functionalities associated with the basic input handler, distribution function module, sampling module, and simulation engine are discussed. Using FD-Gen, synthetic data for heat detector temperature are obtained for 4800 cases in a single-story building with three compartments. As a demonstration, these data are used to train two simple machine learning algorithms. Preliminary results shown that the proposed models can predict the relative location of the fire with reasonable accuracy. Real-time predictions using these methods may be used to help to develop a system to enhance situational awareness for fire fighting.

[4] Accuracy is defined as the number of correct classified instances over the total number of instances.

Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

References

[1] Hamins, A.P., Bryner, N.P., Jones, A.W. and Koepke, G.H., 2015. Research Roadmap for Smart Fire Fighting (No. Special Publication (NIST SP)-1191).

[2] Qolomany, B., Al-Fuqaha, A., Gupta, A., Benhaddou, D., Alwajidi, S., Qadir, J. and Fong, A.C., 2019. Machine Learning, Big Data, And Smart Buildings: A Comprehensive Survey. *arXiv preprint arXiv:1904.01460*.

[3] Mahdavinejad, Mohammad Saeid, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, and Amit P. Sheth. "Machine learning for Internet of Things data analysis: A survey." *Digital Communications and Networks* 4, no. 3 (2018): 161-175.

[4] Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

[5] Chenebert, A., Breckon, T.P. and Gaszczak, A., 2011, September. A non-temporal texture driven approach to real-time fire detection. In *2011 18th IEEE International Conference on Image Processing* (pp. 1741-1744). IEEE.

[6] Yin, Z., Wan, B., Yuan, F., Xia, X. and Shi, J., 2017. A deep normalization and convolutional neural network for image smoke detection. *Ieee Access*, 5, pp.18429-18438.

[7] Shen, D., Chen, X., Nguyen, M. and Yan, W.Q., 2018, April. Flame detection using deep learning. In *2018 4th International Conference on Control, Automation and Robotics* (ICCAR) (pp. 416-420). IEEE.

[8] Aslan, S., Güdükbay, U., Töreyin, B.U. and Çetin, A.E., 2019. Deep Convolutional Generative Adversarial Networks Based Flame Detection in Video. *arXiv preprint arXiv*:1902.01824.

[9] Zhao, Y., Ma, J., Li, X. and Zhang, J., 2018. Saliency detection and deep learning-based wildfire identification in UAV imagery. *Sensors*, 18(3), p.712.

[10] Celik, T., Demirel, H., Ozkaramanli, H. and Uyguroglu, M., 2007. Fire detection using statistical color model in video sequences. *Journal of Visual Communication and Image Representation*, 18(2), pp.176-185.

[11] McGrattan, K., Hostikka, S., McDermott, R., Floyd, J., Weinschenk, C. and Overholt, K., 2013. Fire dynamics simulator user's guide. *NIST special publication*, 1019(6).

[12] Sharma, J., Granmo, O.C., Goodwin, M. and Fidje, J.T., 2017, August. Deep convolutional neural networks for fire detection in images. In *International Conference on Engineering Applications of Neural Networks* (pp. 183-193). Springer, Cham.

[13] https://www.data.gov/ [Online; accessed 2019-08-26].

[14] Bruns, M.C., 2018. Estimating the flashover probability of residential fires using Monte Carlo simulations of the MQH correlation. *Fire technology*, 54(1), pp.187-210.

[15] https://www.eia.gov/consumption/commercial/data/2012/ [Online; accessed 2019-08-26].

[16] Jones, E., Oliphant, E., and Peterson, P., 2001. SciPy: Open Source Scientific Tools for Python, http://www.scipy.org/ [Online; accessed 2019-08-26].

[17] Peacock, R.D., Reneke, P.A. and Forney, G.P., 2017. CFAST–Consolidated Model of Fire Growth and Smoke Transport (Version 7) Volume 2: User's Guide. *NIST Technical Note 1889v2*.

[18] Schittenkopf, C., Deco, G. and Brauer, W., 1997. Two strategies to avoid overfitting in feedforward networks. *Neural networks*, 10(3), pp.505-516.

[19] Peacock, R.D., McGrattan, K.B., Forney, G.P. and Reneke, P.A., 2017. CFAST—Consolidated Fire and Smoke Transport (Version 7)—Volume 4: Verification and Validation Guide. *NIST Technical Note 1889v4* (National Institute of Standards and Technology, Gaithersburg, MD, 2017).

Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.

Appendix A: Summary of important experimental parameters [19].

Test Series	$\dot{Q}$ (kW)	D (m)	H (m)	$\dot{Q}^*$	L_f/H	ϕ	W/H	L/H	r_{cj}/H	r_{nd}/D
ATF Corridors	50 – 500	0.5	2.4	0.3 – 3.3	0.3 – 0.9	0.0 – 0.1	0.8	7.1	0.8 – 6.0	N/A
Fleury Heat Flux	100 – 300	0.3 – 0.6	Open	0.3 – 5.5	Open	Open	Open	Open	Open	1.7 – 3.3
FM/SNL	470 – 516	0.9	6.1	0.6 – 2.4	0.3 – 0.6	0.0 – 0.2	2.0	3.0	0.2 – 0.3	N/A
iBMB*	3500, 400	1.13, 0.79	5.7, 5.6	2.4, 0.7	0.6, 0.4	0.6, 0.1	0.6	0.6		N/A
LLNL Enclosure	50 – 400	0.6	4.5	0.2 – 1.5	0.1 – 0.4	0.1 – 0.4	0.9	1.3	0.3 – 1.0	N/A
NBS Multi-Room	110	0.3	2.4	1.5	0.5		1.0	5.1		N/A
NBS Single-Compartment	2900 – 7000	1.1 – 1.7	2.4	1.7 – 2.3	1.1		1.0	1.5		N/A
NIST Seven-Story	1130	0.7	2.6	2.2	1.1		0.7	5.6		N/A
NIST/NRC	350 – 2200	1.0	3.8	0.3 – 2.0	0.3 – 1.0	0.0 – 0.3	1.9	5.7	0.3 – 2.1	2.0 – 4.0
NIST/NRC Cabinet	200 – 400	0.3 – 0.5	2.1	0.3 – 3.7	0.2 – 0.9	1.3 – 12	0.3	0.4	N/A	1.2 – 2.0
NIST/NRC Corner	200 – 400	0.7	3.8	0.4 – 0.9	0.3 – 0.5	<0.1	1.8	2.9	0.5 – 2.3	N/A
NIST Smoke Alarm	100 – 350	1.0	2.4	0.1 – 0.3	0.2 – 0.5		1.7	8.3	1.3 – 8.3	N/A
PRISME	480 – 1600	0.7 – 1.1	4.0	1.1	0.5 – 0.8	0.5	1.3	1.5	0.0 – 0.5	2.3 – 5.7
SP AST	450	0.3	2.4	6.1	1.1	0.1	1.0	1.5		N/A
Steckler	31.6 – 158	0.3	2.1	0.8 – 3.8	0.3 – 0.7	0.0 – 0.6	1.3	1.3		N/A
UL/NFPRF	4400 – 10000	1.0	7.6	4.0 – 9.1	0.7 – 1.0	Open	4.9	4.9	0.6 – 3.9	N/A
UL/NIST Vents	500 – 2000	0.9	2.4	0.7 – 2.6	0.8 – 1.6	0.2 – 0.6	1.8	2.5	1.0 – 2.3	
USN Hawaii	100 – 7700	0.3 – 2.5	15	0.7 – 1.3	0.1 – 0.4	Open	4.9	6.5	0 – 1.2	N/A
USN Iceland	100 – 15700	0.3 – 3.4	22	0.7 – 1.3	0.0 – 0.3	Open	2.1	3.4	0 – 1.0	N/A
Vettori Flat	1055	0.7	2.6	2.5	1.1	0.3	2.1	3.5	0.8 – 2.9	N/A
VTT Large Hall	1860 – 3640	1.4 – 1.8	19	0.7	0.2	0	1.0	1.4	0 – 0.6	N/A
WTC	1970 – 3240	1.6	3.8	0.6 – 0.9	0.8 – 1.1	0.3 – 0.5	0.9	1.8	0.0 – 0.8	0.3 – 1.3

Appendix B: Summary of statistics for all quantities of interest.

Quantity	σ_E	σ_M	δ
HGL Temperature	0.07	0.32	1.09
HGL Temperature: Forced Ventilation	0.07	0.20	1.13
HGL Temperature: Natural Ventilation	0.07	0.39	1.08
HGL Temperature: No Ventilation	0.07	0.12	0.93
HGL Depth	0.05	0.27	1.01
HGL Depth: Open Compartments	0.05	0.17	0.94
HGL Depth: Closed Compartments	0.05	0.24	1.45
Ceiling Jet Temperature	0.07	0.45	1.02
Plume Temperature	0.07	0.23	1.08
Oxygen Concentration	0.08	0.26	1.03
Carbon Dioxide Concentration	0.08	0.27	0.89
Carbon Monoxide Concentration	0.19	0.65	1.04
Smoke Concentration	0.19	0.68	3.43
Compartment Over-Pressure	0.23	0.56	1.45
Target Temperature	0.07	0.50	1.25
Surface Temperature	0.07	0.21	1.00
Target Heat Flux	0.11	0.60	0.97
Surface Heat Flux	0.11	0.23	0.91
Smoke Alarm Activation Time (Temperature Surrogate)	0.34	0.43	1.23
Smoke Alarm Activation Time (Smoke Obscuration)	0.34	0.51	0.56
Sprinkler Activation Time	0.06	0.20	1.01

Note that the term δ is a calculated bias factor representing the degree to which the model over-predicted or under-predicted experimental data, the term σ_M is a measure of model uncertainty, and the term σ_E is a measure of experimental uncertainty. The expression $\delta > 0$ means the model over-predicted the observations, and $\sigma_M < \sigma_E$ means that the model uncertainty is within experimental uncertainty.

Tam, Wai Cheong; Cleary, Thomas; Fu, Eugene Yujun.
"Generating Synthetic Sensor Data to Facilitate Machine Learning Paradigm for Prediction of Building Fire Hazard."
Paper presented at Suppression, Detection and Signaling Research and Applications (SUPDET 2019), Denver, CO, United States. September 17, 2019 - September 20, 2019.