Journal of Research of the National Bureau of Standards
Volume 90, Number 6, November-December 1985

The Limitations of Models and Measurements as Revealed Through Chemometric Intercomparison

L.A. Currie
National Bureau of Standards, Gaithersburg, MD 20899

Accepted: July 1, 1985

Interlaboratory Comparisons using common (reference) materials of known composition are an established means for assessing overall measurement precision and accuracy. Intercomparisons based on common data sets are equally important and informative, when one is dealing with complex chemical patterns or spectra requiring significant numerical modeling and manipulation for component identification and quantification. Two case studies of "Chemometric Intercomparison" using Simulation Test Data (STD) are presented, the one comprising STD vectors as applied to nuclear spectrometry, and the other, STD data matrices as applied to aerosol source apportionment. Generic information gained from these two exercises includes: a) the requisites for a successful STD intercomparison (including the nature and preparation of the simulation test patterns); b) surprising degrees of bias and imprecision associated with the data evaluation process, per se; c) the need for increased attention to implicit assumptions and adequate statements of uncertainty; and d) the importance of STD beyond the Intercomparison—i.e., their value as a chemometric research tool. Open research questions developed from the STD exercises are highlighted, especially the opportunity to explore "Scientific Intuition" which is essential for the solution of the underdetermined, multicollinear inverse problems that characterize modern Analytical Chemistry.

Key words: aerosol source apportionment; chemometric intercomparison; gamma-ray spectra; interlaboratory comparison; inverse problem; linear regression; multivariate data analysis; pattern recognition; reference materials; scientific intuition; scientific judgment; simulation test data.

Introduction
Accuracy Assessment

Ideally, the results of chemical analyses performed by a single laboratory using a well-defined Chemical Measurement Process (CMP) should be characterized by reliable measures of accuracy—i.e., imprecision and bias (or bounds for bias). Meaningful statements of uncertainty would then follow directly from these CMP Performance Characteristics [1][1]. Such is almost never the case, however. Once two or more laboratories perform measurements of the same material, interlaboratory errors become evident. Collaborative tests, using common, homogeneous materials, serve as one of the most powerful means for both exposing and estimating the magnitude of this error component.

A familiar illustration of the outcome of interlaboratory measurement is reproduced in figure 1 [2]. Here, following the spirit of the "Youden Plot" [3], we show the results of pairs of measurements by 10 laboratories of the determination of trace levels of vanadium in two Standard Reference Materials (SRMs). A log transform has been applied to the reported concentrations, in order to expose proportionate errors among laboratories. As is generally the case, intralaboratory precision is comparable among laboratories, and considerably better than the interlaboratory component. Note that the line drawn in the figure is *not* fitted; its location is fixed by the certified values of the SRMs (dashed box),

About the Author: L.A. Currie is with the NBS Center for Analytical Chemistry where he leads the atmospheric chemistry group.

[1]Figures in brackets indicate literature references.

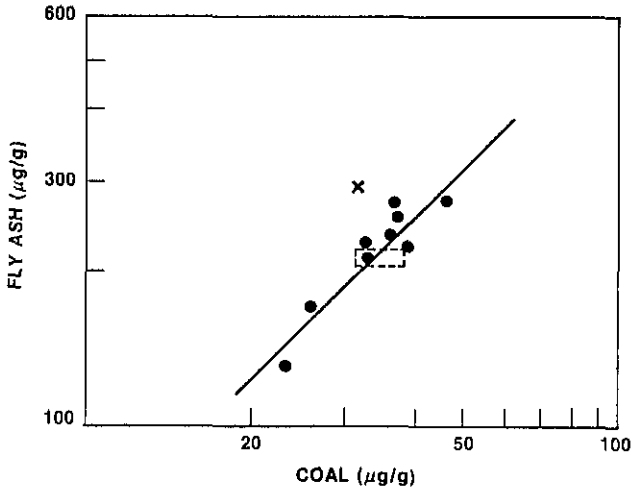

Figure 1—Interlaboratory results for vanadium (μg/g) in two standard reference materials. The plot shows proportionate interlaboratory errors among nine participants using the same analytical method. The outlier (*x*) derived from a second method, lacking internal replication. Dashed region indicates the 'truth' [certified values].

and its slope is fixed at 45° (proportionate errors). SRMs or samples having known composition carry a very important attribute in interlaboratory tests, in that one can estimate individual laboratory and CMP bias in addition to interlaboratory variability. Also noteworthy is the "outlier" (marked by the cross) which deviates from the line significantly more than the other members of the interlaboratory set. Investigation of such outliers can sometimes yield important insight into the causes of interlaboratory differences. (In the example at hand the "outlier" resulted from a different analytical method that lacked internal replication.)

Modern Analytical Chemistry: Importance of Data Evaluation

Enormous advances in analytical methods have brought great improvements in sensitivity, but at the same time, significant complications in data interpretation. In little over a decade, for example, "trace analysis" came to mean the measurement of pg of an analyte rather than μg [4]. Chemical patterns or spectra are central to the interpretation of complex mixtures, as are sophisticated analyte separation techniques such as high resolution gas chromatography. Practical demands on analysts also have accompanied the increase in sensitivity; toxic chemicals, for example, are regulated down to concentrations of 10^{-12} g/g. The magnitude of the problem can be appreciated from the fact that the analyst in measuring a substance at a concentation of 10^{-11} g/g in drinking water must contend with $\sim 10^5$ compounds which are more concentrated by at least a factor of 1000 [5].

When the Chemical Measurement Process involves significant modeling or numerical operations in the data evaluation or information extraction step, it becomes interesting to consider the data analog of the SRM—the STD or Simulation Test Data set. By providing participants with common, well-characterized sets of data which adequately simulate the observations of real experiments, one can directly assess the imprecision and bias of the data evaluation process, independent of confounding errors or unreliable assumptions connected with the experimental parts of the CMP. This, in turn, makes it possible to estimate the error components associated purely with the experimental steps. Simulaton Data, as opposed to Real Data, are beneficial because "the truth is known"—i.e., the physical model (functional relation) as well as the random error model can be strictly controlled.

One might expect that little may be learned from such "Chemometric Intercomparisons" since numerical operations can be reproduced quite rigorously from laboratory to laboratory; but such is not the case. An illustration involving Real Data comes from the reevaluation (auditing) of several sets of chromatographic data from Love Canal soil and sediment samples for toxic organic compounds. As shown in table 1, compounds identified in common between analytical and auditing labs represented only about 60% of the total identifications, where the discrepancy was due strictly to differences in data evaluation [6].

Table 1. Real data—Love Canal soil and sediment samples: Compound identification by GC/MS. (Data Tape Auditing).

Lab Code	Same Compounds Identified	Total Compounds Identified
A	20	32
B	13	24
C	22	51
D	13	20
E	63	104
EPA	14	20
	[intersection]	[union]

A myriad of hidden assumptions, and even algorithm changes exist in many of the pattern recognition and spectrum deconvolution schemes currently in vogue. Since the actual number of degrees of freedom is generally negative—i.e., the chemical model is never really known—numerical solutions often require subtle injections of "scientific intuition" or "scientific judgment." The importance of these issues will be illustrated by two case studies, actual intercomparisons among expert laboratories of the data evaluation phases in Gamma Ray Spectrometry and trace element Aerosol Source Apportionment, respectively. The STD in the first exercise was a data vector (nuclear spectrum); in the second, it was a data matrix (set of samples each having a trace element "spectrum"). The author was a participant in the first intercomparison and instigator of the second.

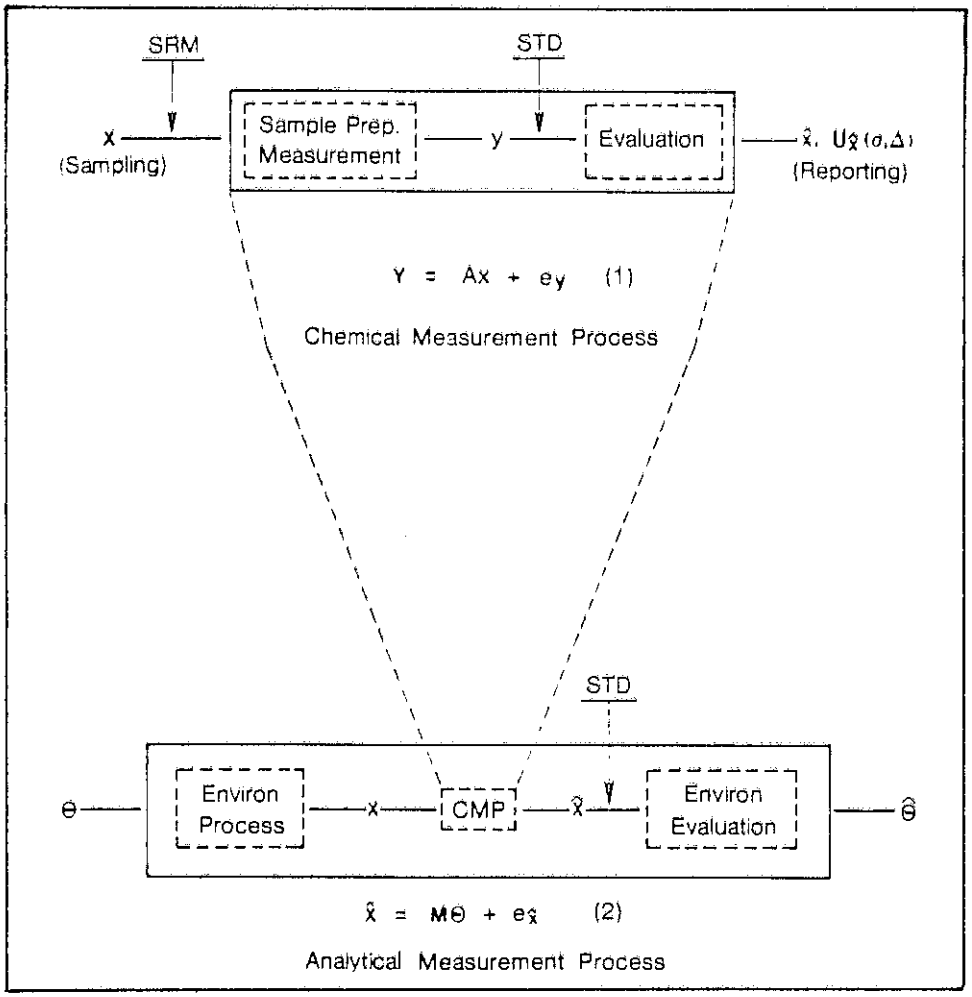

Figure 2—The Chemical Measurement Process (CMP) which operates within the laboratory, and the Analytical Measurement Process (AMP) which operates within the larger "Environmental" (or other external) system. (See text for explanation of symbols.)

Chemometric Intercomparison

Structure of the Measurement Process

In order to introduce some notation and to put the STD Intercomparison (IC) in perspective, it is useful to consider the structure of the Chemical and Analytical Measurement Processes (CMP, AMP) [4]. These two processes, which symbolize the environment in which Analytical Chemistry operates, are shown in figure 2. As indicated in the upper portion of the figure, the CMP represents the laboratory process, where a sample of composition **x** is operated on (chemically) to produce a signal **y**, which in turn is operated on (mathematically) to generate an estimate of **x** and an uncertainty interval. The chemometric challenge thus is to obtain a chemically-meaningful and mathematically-consistent solution to the inverse problem as represented by eq 1. Control of the overall measurement process is achieved by injection of an SRM as a surrogate sample; control of the data evaluation process is achieved by injection of an STD as a surrogate signal.

Except in limited laboratory investigations, the real object of Analytical Chemistry is to provide information on an external attribute (here represented by $\boldsymbol{\theta}$) through compositional analysis. The lower portion of figure 2 describes this broader context, where an external process (here labeled "environmental") operates an $\boldsymbol{\theta}$ to produce the sample of composition **x**. Following this, the imbedded CMP yields the compositional estimate $\hat{\mathbf{x}}$. The final step, once again, is the solution of a (generally more difficult) inverse problem eq 2. STD injection in the AMP case means provision of a surrogate sample whose estimated compositional pattern corresponds to eq 2.

A fundamental difference exists between the CMP and the AMP with respect to the chemometric task. That is, in the laboratory, in principle, we can isolate ever-decreasing numbers of analytes (chemical fractions or instrumental signatures), in many cases leaving just a single term (component) in eq 1. For the AMP and the corresponding environmental, geochemical, or biochemical problem, for example, Nature is seldom so cooperative. That is, real samples **x** are determined by the external process over which we have limited control (beyond the sampling design), so eq 2 nearly always exhibits a multicomponent, multivariate structure. Unique solutions are generally impossible in the absence of

scientific knowledge concerning the external ("environmental") system.

The following STD intercomparison consists of univariate data ($\mathbf{y}$) from a simulated CMP. The second example consists of multivariate data ($\hat{\mathbf{x}}$) from a simulated AMP. Both ICs took place because of analytical measurement problems having major public import—the first related to accurate monitoring of radioactivity; the second, to accurate apportionment of atmospheric pollutants.

STD Vector— IAEA Intercomparison of Gamma Ray Spectrum Analyses

In connection with their Analytical Quality Control Services program, the International Atomic Energy Agency (IAEA) undertook in 1976-77 a broad interlaboratory data evaluation exercise involving computer-simulated high resolution Ge(Li) gamma ray spectra such as might arise in contemporary neutron activation analysis [7]. The purpose of the intercomparison was both to assess the state of the γ-spectrum evaluation art and to provide data sets of known structure to assist in the improvement of that "art" (or science?). To my knowledge this was the first numerical chemometric intercomparison of such scope, using STD vectors. The organization and structure of the IAEA exercise are summarized in table 2.

Several features of the IAEA intercomparison were analogous to those involving chemical measurement intercomparisons with reference materials. First, the STD were well-characterized, with γ-rays of known identity (energy) and amplitude. The "samples" were absolutely homogenous (identical numerical data to all participants), and they simulated observations from actual laboratory samples. SRM intercomparison organizers strive to also meet such conditions, but of course they can only approach the homogeneity and exact composition knowledge found with STD. The IAEA data sets also had known random error distributions (Poisson), a situation which is actually approached in many nuclear experiments, but which can never be guaranteed. Realism was preserved in the shapes of the γ-peaks, in that they were derived from high precision observations with Ge(Li) spectrometers. The fact that these shapes were not analytic was one of the more discriminating elements of the IC, particularly for the resolution of doublets, where alternative analytic or empirical peak shape functions *had* to be employed [8]. (Each peak was approximately Gaussian near the top but decidely asymmetric near its base.) Referring again to table 2, we can see that important planning took place, over a three-year period, resulting in four categories of data designed to provide initial "calibration," and to test detection, accuracy and precision in quantification, and doublet resolution. The importance of the pilot study cannot be overstated; development of realistic STD of sufficient but not excessive complexity does not come about without careful initial trials and iteration.

Detailed results for the IC may be found in [7]. Some of the highlights follow. Figure 3, for example, shows the spectrum (pattern) offered the participants, in digital and analog form, to address the problem of unknown peak detection. Participants knew only the calibration peak shape (as a function of "energy" or channel number) from the Reference Spectrum #100 (not shown), plus the facts that the unknown peaks were singlets and that random errors were Poisson. The numbers and locations of the trace peaks were to be determined. (The steep rise in the baseline near the center of the spectrum was inserted by the IAEA to simulate a Compton Edge.) The inset shown in figure 3 gives some

Table 2. Structure of the IAEA Gamma-Ray STD intercomparison.

Objectives
• To permit each participant to assess the accuracy of his data evaluation process.
• To determine the quality of alternative gamma-ray spectrum evaluation methods as applied in representative laboratories.
Evolution
• 1973: Proposed at Consultants' Meeting.
• 1975-6: Pilot Study involving a small number of experts.
• 1976-7: Full IC, involving 163 labs in 34 member states.
• Currently: Simulation data offered as continuing part of the IAEA Analytical Quality Control Service.
Data Sets
• Reference Spectrum: 20 high-precision peaks spanning 2000 channels.
• Detection Spectrum: 22 subliminal peaks, whose number and locations were unknown to participants; detection criteria (α-, β-errors) were left to individual judgment.
• Precision Spectra: 6 replicate spectra having 20 known plus 2 unknown, large singlet peaks (Poisson statistics).
• Resolution Spectrum: 9 doublets of unknown location and relative amplitude.

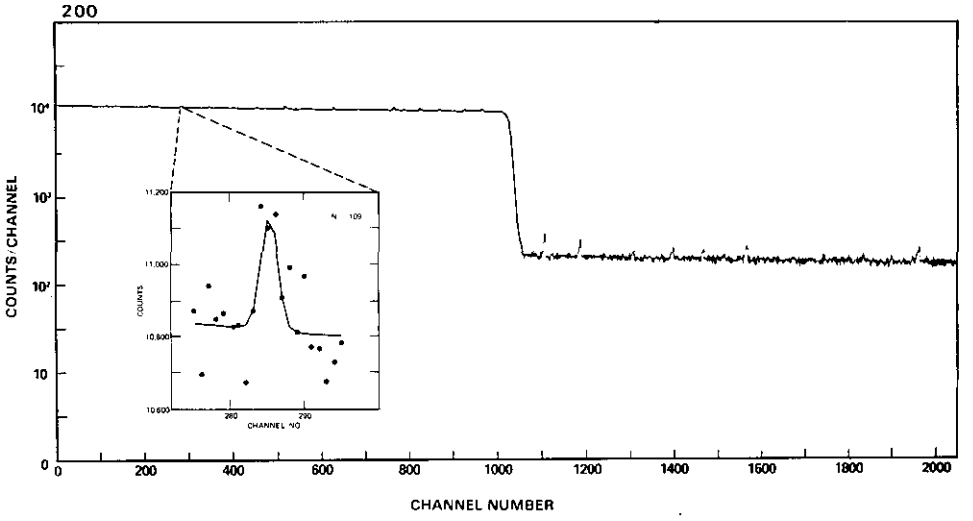

Figure 3—IAEA Gamma-Ray STD; "Detection" Spectrum. Inset shows discrete data for a peak detected by about 50% of the participants. Adapted from [7.]

idea of the discreteness and scatter of the digital data; this "real" peak was detected by about half of the participants.

The results of this intercomparison were somewhat surprising. Though most of the 200-odd participants submitted results, no one correctly identified all 22 subliminal peaks. Some six classes of methods, including one labeled "unclassified," were employed, including: the relative maximum, first and second derivatives, cross correlation, and "visual". It is interesting that the last gave the best result; one "trained eye," using analog data only, identified 19 peaks correctly, with no false positives! Understanding the process (Scientific Intuition) employed by this expert is certainly one of the more intriguing aspects of this work. It seems hardly pure chance, for only 5 out of 212 participants correctly reported this many (19) peaks; yet 2 of the 5 were "visual." (For comparison, the visual technique was employed by about 5% of the participants.) The second derivative and cross correlation techniques were close behind, with up to 18 and 17 peaks correctly identified (w/o false positives), respectively. Performance was quite diverse for all methods, however: correct identifications ranged from 2 to 19 peaks, and false positives ranged from 0 to 23. Apparently, Detection Limits were rarely estimated, for this issue was not even mentioned in [7]. Histograms for the three "best" methods are shown in figure 4. Though all three exhibit considerable dispersion, it is clear that the visual technique gave the best single result as well as the smallest fraction of false positives.[2]

The replications (Spectra 300-5) and resolution (Spectrum 400) exercises also indicated often inadequate and

SCIENTIFIC INTUITION

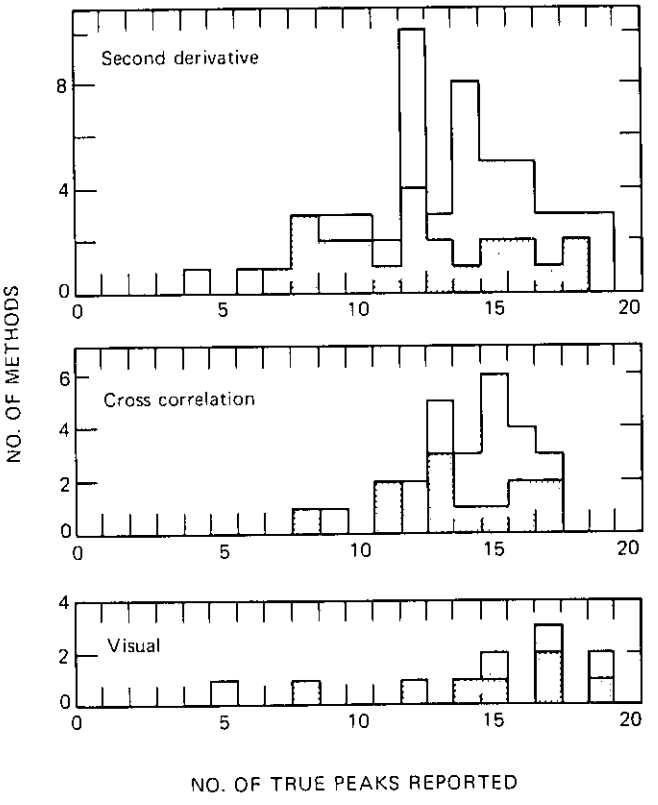

Figure 4—Frequency distribution of results for peak detection according to the type of method used. The upper boundary of each histogram represents the data for all results regardless of the number of spurious peaks reported; the upper boundary of the shaded region is for those results which were accompanied by zero spurious peaks.

[2]Scientific Intuition, as employed by experts, is alleged to be much more disperse that "rule-based" methods [9]. In the light of figure 4, it is not obvious that this presumption is true, for even the "objective" numerical techniques employed by different laboratories operating on exactly the same data gave broad distributions.

widely varying performance. The majority of the results submitted contained quite inaccurate or no estimates of uncertainty for large singlet peaks, even though the random error distribution was known; and less than 25% of the participants even submitted results for the most difficult doublet resolution case.

The IAEA Simulation Data Sets have been viewed of sufficient importance that they have become an integral part of the Intercomparison Programme of the Analytical Quality Control Service of that organization. The most recent offering was issued in December 1984 [10] where STD γ-ray spectra are included alongside isotopic and trace element Intercomparison and Certified Reference Materials of importance in many areas of nuclear and environmental analysis.

STD Matrix—NBS-EPA Intercomparison of Source Apportionment Techniques. The second case study comprises STD in the form of two-dimensional data matrices, simulating sets of atmospheric aerosol samples each analyzed for up to 20 chemical species [11]. The stimulus for this exercise, which is believed to be the first STD intercomparison involving Data Matrices, was the great potential but great difficulty of identifying multiple pollutant sources via their "chemical fingerprints" as preserved in ambient particles. (A vivid illustration adjoins [11], where one finds discord even in assigning names to pollutant factors deduced from elemental patterns observed in actual measurements of [Houston] aerosol samples [12].) As noted at the beginning of this section, this type of problem is characteristic of the AMP, where superposition of multiple components is intrinsic to the nature of the system, so chemical manipulation cannot simplify the structure of eq (2).

We designed the STD in coordination with (nearly) all of the "Receptor Modeling" (source apportionment) experts in the U.S. with the object of providing a few realistic data matrices covering a range of problems. The overall structure of the study is given in table 3. The data matrix $\mathbf{x}$ is given by the superposition of source contributions $(\mathbf{M\theta})_j$, where each source has a characteristic chemical pattern or profile $\mathbf{M}_j$ and a temporal (or spatial) intensity pattern $\boldsymbol{\theta}_j$. Three classes of error typify such measurements, as indicated in the table. A significant task involved building the database of source profiles and error terms. Unlike the γ-ray calibration profiles (peak shapes), the aerosol source profiles were not even approximately analytic (fig. 5, top); reliable empirical field data had to be sought and evaluated.

When generating data matrix STDs, one must pay attention to a major new element of complexity which is absent from data vector STDs. That is, unless the simulation data are to be no more than arbitrary superpositions of sources with added random errors, it is essential to generate the source intensity patterns by means of suitable source emission locations (emissions inventory map) plus realistic mixing and transport to the receptor (sampling) site(s). This step is intrinsic to the nature of the data matrix; it underlies its multivariate character, and it highlights information (source map, meteorological patterns,...) external to the data matrix, per se, which may be crucial for a successful data analysis. The source map used for the simplest of our three data sets is shown in the lower portion of figure 5. Stochastic source impacts at the receptor $\mathbf{R}$ were generated by bringing together the source map, emissions intensities and operating schedules, actual meteorological data (from St. Louis, September 1976), and the 'RAM' atmospheric dis-

Table 3. Structure of the source apportionment simulation test data.

Generating equation

$$\hat{\mathbf{x}}_{it} = \sum^{P} [\mathbf{M} - \mathbf{e}_m + \mathbf{e}_H]_{ij} \boldsymbol{\theta}_{jt} + \mathbf{e}_{it}$$

where:

p = number of active sources ($p \leq 13$)
t = sampling period ($1 \leq t \leq 40$)
$\hat{\mathbf{x}}_{it}$ = "observed" concentration of species i for period t ($1 \leq i \leq N, N \leq 20$)
$\boldsymbol{\theta}_{jt}$ = true intensity (at receptor) of source j ($1 \leq j \leq p$)
$\mathbf{M}_{ij}$ = 'observed' source profile matrix (element i,j)
$\mathbf{e}_t$ = random measurement errors, independent and normally distributed
$\mathbf{e}_m$ = systematic source profile errors, independent and normally distributed (systematic because fixed over the 40 sampling periods)
$\mathbf{e}_H$ = random source profile variation errors, independent and log-normally distributed

Data set characteristics

Set I: $p = 9$ (including one unknown source);* errors = e_t, e_m; City Plan No. 1 (fig. 5)
Set II: $p = 13$ (all known); errors = e_t, e_m; City Plan No. 2
Set III: $p = 13$ (all known); errors = e_t, e_m, e_H; City Plan No. 2

*For Data Set I, participants were told only that $p \leq 13$.

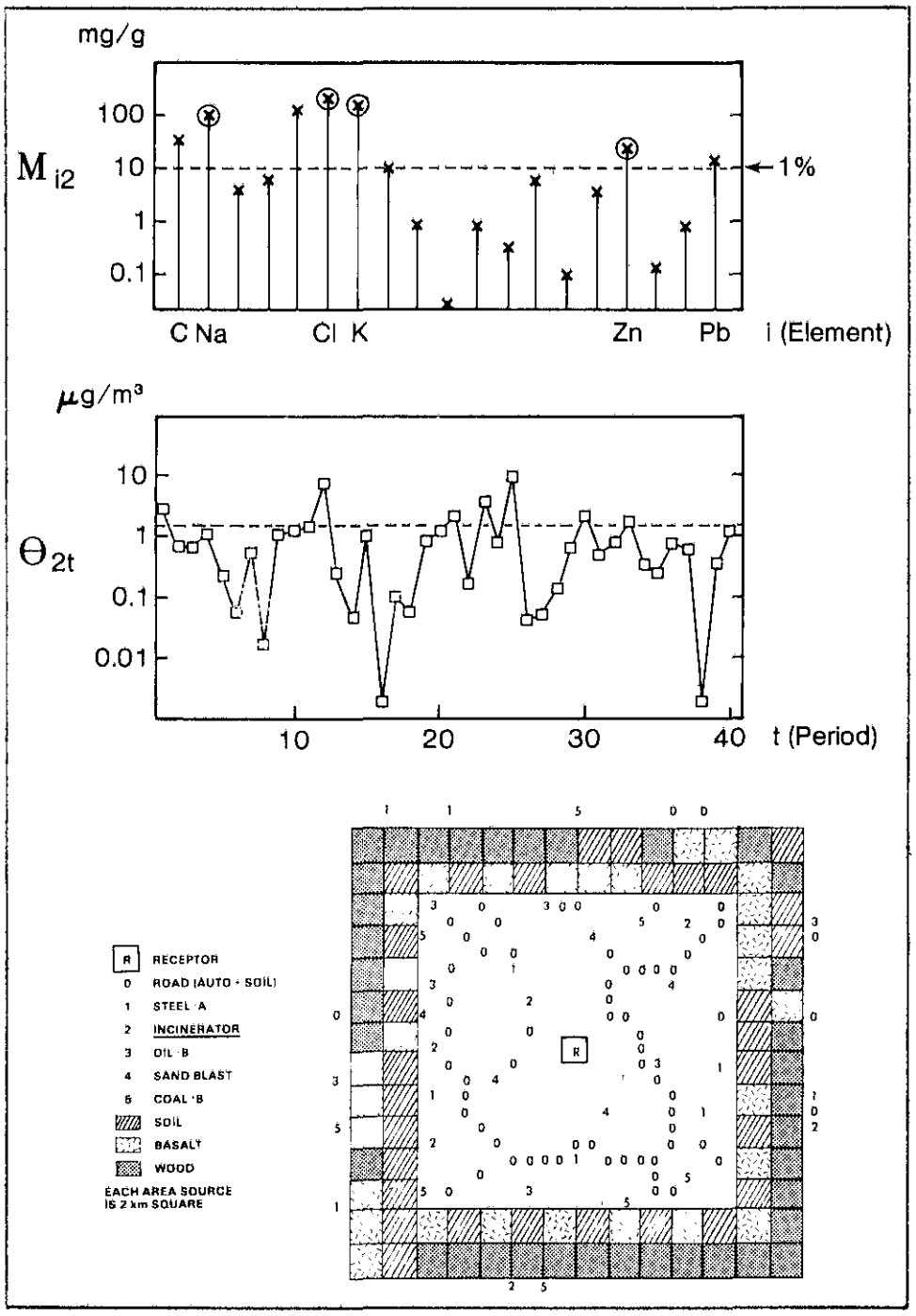

Figure 5—Source Apportionment STD (Data Set I), Upper portion shows one column (transposed) of the source signature matrix **M**, and one row of the source intensity matrix **θ**—both for source-2, INCINERATOR. M_{i2} has a discrete pattern (individual chemical elements), the most discriminating elements of which are marked by circles; the dashed line indicates which elements exceed 1% of the (Incinerator) particle mass. θ_{2t} has a continuous underlying structure (time series) which is sampled at 40 equidistant points; the dashed line indicates samples for which the Incinerator source contributes more than 5% of the average aerosol mass.

The lower portion of the figure displays the aerosol source emission map.

persion model [13]. Illustrations of a source (chemical) signature $[M_{ij}]$ and a source intensity time series $[\theta_{jt}]$ are given in the upper portion of the figure. Note that M_{ij} by its nature (individual elements) is discrete, whereas θ_{jt} is a sampled continous time series (intensity variations).

The objective of spanning the range of difficulty was achieved. Our initial "pilot" data matrix was so transparent that one of our participant-advisors was able to identify sources by inspection. Caused in part by the narrowness of the RAM model plumes, this was remedied in the final STD intercomparison data sets, one of which was so difficult (though realistic) that none of the participants submitted results. Results for the data set of intermediate difficulty (Set II) were generated by three laboratories, all using regression techniques. Two of these were identical: "effective variance" weighted least squares (WLS), which took into account errors in the observed chemical concentrations as well as those in the source profiles. The third method was

ridge regression, of interest because of the high degree of collinearity among the 13 sources. For the most part, the three sets of results were self consistent, and within a factor of 2 of the truth. For four of the sources, however, widely discrepant results were reported—a surprising outcome, for similar or identical numerical methods were applied to identical numerical data.

An exploratory graphical analysis of the results from the two laboratories using weighted least squares (WLS-1, -2) is shown in figure 6. This approach, which was inspired by the "Youden Diagram" for bi-material intercomparisons, allows us to spot "outliers" from the line of concordance. (As with the line drawn for the Youden plot of figure 1, the 45° line in figure 6 was drawn independent of the data—i.e., it was in no way "fitted".) Exact results would fall at the origin (0, 0), and equivalent data reduction by the two laboratories would produce results lying on the line. Dispersion along the line derives from random errors built into the common data set and systematic errors connected with the common numerical model (WLS). Two of the sources (Ag, O_2) were well below their detection limits, and will not be further discussed here. The three outliers (*cf* point 'x' in figure 1)—G, W, and C_1—deserve our attention, however. As will be shown in the next section, they may be attributed to what the laboratories involved labeled "Scientific Judgment." Discovering this factor, and understanding its nature, was one of the unexpected but important outcomes of the experiment.

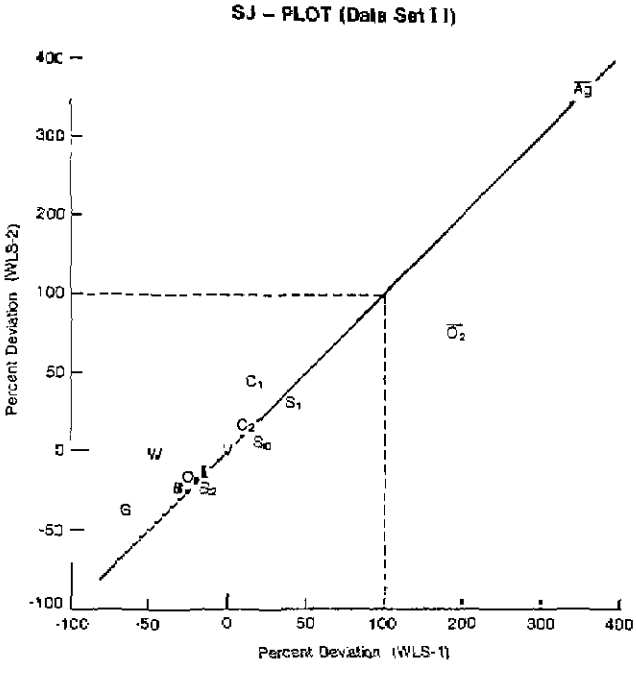

Figure 6—"Scientific Judgment" Plot. Correlation (Youden-type) Diagram showing data evaluation results of laboratory-2 vs laboratory-1 operating on the same data set (II) using the same method of numerical analysis (Weighted Least Squares). Deviations from the *non-fitted* line of concordance imply chemometric "operator error" (SJ).

[Source Codes: Steel-A (S_1), Steel-B (S_2), Oil-A (O_1), Oil-B (O_2), Incinerator (I), Glass Mfr (G), Coal-A (C_1), Coal-B (C_2), Aggregate (Ag), Basalt (B), Soil (So), Auto (V), Wood Smoke (W).]

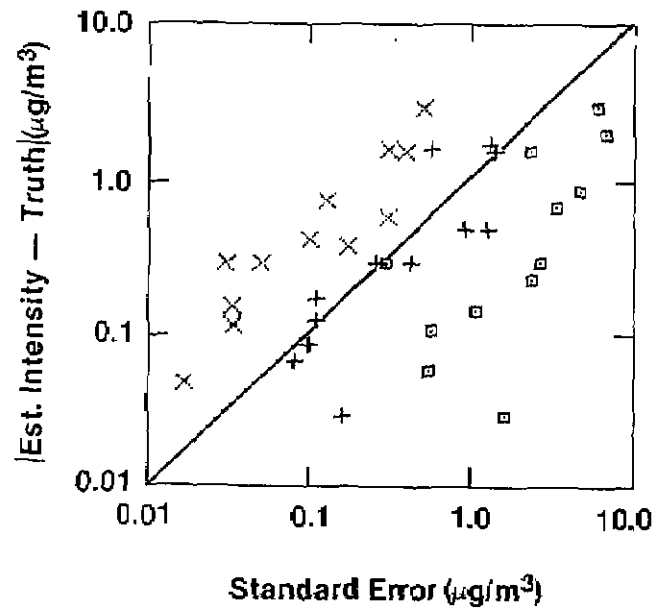

Figure 7—Absolute errors for each of the 13 source estimates for Data Set II, by WLS-1 (×), WLS-2 (+) and Ridge Regression (□), plotted as a function of the reported standard errors. (On the average about 2/3 of the points for a correct method should fall below the diagonal.)

Another major difference and inadequacy among the laboratories relates to the uncertainties (i.e., SEs) reported. All did report standard errors, but an examination of the actual deviations from the truth for the 13 estimates from each lab was illuminating. For one of the three, all 13 (absolute) deviations exceeded the SEs by factors of 2 to 10, whereas for another lab the deviations were all smaller than the SEs by factors of about 1.6 to 30. None of the labs reported bounds for systematic or model error. (See fig. 7.)

Because of the multivariate character of the source apportionment data sets, both simple linear regression and factor analysis (FA) techniques could be applied. The regression methods ("Chemical Mass Balance", i.e., WLS) were the more precise, when model information and source profiles were available. When the model was not fully known and if the number of interfering components was not too great, factor analysis or certain hybrid methods appeared to yield more acceptable results. Such was the case for Data Set I which had 9 sources, one of which was unknown to the participants. In Set II, however, with 13 known sources having a significant degree of multicollinearity, factor analysis was able to discern but four sources.

Subsequent Developments: STD as a Basis for Chemometrics Research

A principal conclusion from the foregoing exercises is that the data evaluation step of multicomponent and multidimensional chemical (nuclear) analysis is *not* free from problems of imprecision and bias. Scientific Intuition (SI) and Scientific Judgment (SJ), often manifest as subtle assumptions, can provide important guidance for pattern recognition, or it can be somewhat misleading. Collaborative STD exercises appear to be an effective means for exposing and perhaps better understanding these "expert" techniques.

The limitations found in the STD intercomparisons—*e.g.*, limitations in accuracy, in model formulation, in uncertainty estimation, in detection, and in utilizing external information (non-negativity, meteorological data,...)—suggest that an important part of experimental inaccuracy, as seen for example in SRM intercomparisons, may lie with the data evaluation process. Attention to this matter offers the possibility of improved overall performance, either through the reduction of needless data evaluation error, or through improved measurement process design to reduce model complexity and multicollinearity.

Both intercomparison exercises exhibited an afterlife. As noted earlier, the Gamma-Ray STD have become a routine part of the IAEA's Analytical Quality Control Service. The source apportionment STD have spontaneously evolved into a research data set for chemical element pattern recognition method development. Initially, new research was undertaken by participants who wished to try improved versions of the their regression or FA methods in light of the IC results and knowledge of the truth. More recently, requests for the data tape have come from others for the testing and development of multivariate pattern recognition methods quite apart from aerosol source apportionment. A sampling of post-intercomparison research with the STD data matrices follows:

Investigator	Topic	Ref.
M-D. Cheng, P. Hopke*	Linear Programming	[14]
L. Currie*	Detection, Design, Model Error	
I. Frank, B. Kowalski	Partial Least Squares	[15]
G. Gordon*	Student Instruction, QA	
R. Henry	Composite Components (SVD)	[16]
P. Lioy	Student Research	
D. Lowenthal et al.	Special Error Propagation (Cov)	[17]
T. Pace	Sensitivity Analysis	

*Participants of the original IC, known to be performing advanced studies using the STD. Most of the others listed subsequently requested the data tape specifically for basic investigations of the numerical data evaluation process.

Table 4 includes results from some of these more recent investigations, together with an examination of the "outliers" and Scientific Judgment exposed in the correlation diagram (fig. 6). Table 4A lists results for the best prior results (TTFA) for Data Set I and the new application of Partial Least Squares (PLS). Though PLS results carry no uncertainty estimates, they are clearly closer to the truth. The two deceptions, as well as all of the source profiles, however, were known for carrying out the PLS analysis. Regarding TTFA and other prior analysis, the first deception—the linear combination of AUTO and SOIL to represent the ROAD component, including re-entrained dust—was missed by all participants; the presence of a large, additional component (SANDBLAST) was discovered by all.[3]

Table 4A. Source apportionment STD (Data Set I).

Source	Truth [μg/m³]	PLS[1]	TTFA[2]	Δ/SE[2]
Steel (c)	0.05	0.07	—	?
Oil (c)	2.0	2.5	2.1±1.0	+0.1
Incinerator	1.3	1.4	1.9±0.14	+4.3
Coal-B	2.4	1.9	2.2±0.73	−0.3
"Crustal" (c)	12.7	13.0	12.5±0.75	−0.3
*Road	7.1	7.2	*4.0±0.09*	*−34.4*
Wood	3.3	3.4	4.3±0.52	+1.9
*Sandblast	4.2	4.1	4.1±0.20	−0.5
(Total)		(+2%)	(−6%)	

(c) Composite components for multicollinearity reduction.
*Two Deceptions: Road (Soil+Auto); Sandblast (Unk).

[1]I. Frank and B. Kowalski [15]. [2]P. Hopke [11].

Table 4B. Source apportionment STD (Data Set II)—analysis of outliers.

	Incinerator (*I*)	Glass (*G*)	Coal-A (C_L)	Wood (*W*)
Truth (μg/m³)	1.81	0.46	3.7	0.94
[*N*]	[38]	[36]	[32]	[40]
WLS-1[1]	1.5±0.05	0.16±0.03	4.3±0.3	0.5±0.1
[*N*]	[27]	[3]	[11]	[11]
WLS-2[2]	1.5±0.43	0.28±0.11	5.3±1.4	0.91±0.16
[*N*]	[26]	[8]	[22]	[26]
WLS-3[3]	1.51±0.14	0.43±0.44	4.3±2.2	1.33±0.24
[*N*]	[31]	[23]	[29]	[36]
L1[4]	1.60	0.48	0.44	0.45

N=Number of actual or assumed non-zero occurrences of the source in question among the 40 aerosol samples.

[1]Heisler, Shah [11]. [3]Lowenthal, Hanumara, Rahn, Currie [17].
[2]Cooper, DeCesar [11]. [4]Cheng, Hopke [14].

The outliers (G, C_1, W) from data Set II are examined in table 4B for new and earlier analyses. The INCINERATOR results, which fell directly on the line of concordance (fig. 6), are included for comparison. It is interesting that L1 linear programming gave significantly improved results for sources I and G, but much poorer results for the other two sources. A conjecture, which bears investigation, is that in the absence of experimental blunders, least absolute residuals carry too high a penalty in highly collinear problems. We gain some understanding of Scientific Judgment (SJ) from the three WLS results. It is apparent that the exercise of SJ in deleting rare components from the model is operator-dependent and generates important differences in overall results. This is an issue deserving additional research, the outcome of which could be the transformation of Ad-Hoc SJ into scientifically-based SI.[4]

Conclusion

For modern Analytical Chemistry, where chemometric approaches are mandatory to resolve complex signals from multianalyte mixtures, Simulation Test Data serve two important purposes: 1) The assessment of interlaboratory data-evaluation precision and accuracy; and 2) the exposure of SI and SJ plus the generation of more powerful methods of combining scientific knowledge with advanced techniques of data analysis. The need for this is critical, because in principle nearly all of our multicomponent "inverse" problems are underdetermined—i.e., solutions cannot obtain in the absence of explicit (or hidden) assumptions. We conclude with a summary of recommendations for the preparation of STD data sets, and open research questions which could be fruitfully addressed by chemometricians (table 5).

Table 5. STD exercises.

Observations and Recommendations

- Benefits: Controlled model, errors; Truth is known.
- Prerequisites: defined objectives, 'lab' population, plan, input data/model.
- Pilot study advisable.
- Investigate 'surprises'—unexpected accuracy, discrepancies.
- Blind; anonymous; don't score; discourage premature communication.
- Deceptions: mimic reality; be not too obvious.

Some Open Questions

- Adequate treatment of uncertainty—esp. bias (bounds).
- Further understanding of 'SJ' and 'SI'—who are the experts?
- Utilization of external information: phenomena, data, "fuzzy" & partial knowledge, constraints,...
- Treatment of the 'total' problem (simultaneous estimation over the entire data matrix), incl. errors in 'y' and 'A'.
- Solution and uncertainties when the linear model doesn't apply (physical-chemical and mathematical strategies; eg, atmospheric transformation).

Special recognition should be given to the designers of the two STD exercises discussed, as well as to the participants and the scientists who are using the STD for continued basic research in chemometrics. R.M. Parr (IAEA - STD) and R.W. Gerlach (Source Apportionment - STD) made essential contributions. Others deserving special recognition are listed as authors of references [7, 11], and [14 - 18].

[3]STD deceptions—*i.e.*, realistic complications—are, of course, in order for all but the simplest of exercises. Though organizers should attempt to span the range of difficulty occurring with real data sets, they should not do so in too obvious or regular a manner. Informing the IAEA γ-ray STD participants, for example, that the multiplets had just 2 components was already a considerable simplification, but the regular spacing (1, 3, 10 channels) and relative amplitude (1, 3, 10) of the doublet members was unnatural and could encourage a certain amount of guess work on the part of the participants.

[4]The paradigm proposed here classifies experts' decisions or assumptions as "intuitive" (SI) or "judgmental" (SJ) depending on whether they are based on sound (though possibly subliminal) reasoning or *ad hoc* judgments, respectively. These asymptotic classes may each yield correct or incorrect results, just as target and contaminating populations may each produce outliers or inliers, though with differing probabilities.

References

[1] Eisenhart, C., Science **160** (1968) 1201.
[2] Currie, L.A., and J.R. DeVoe, Chapt 3 in J.R. DeVoe, Ed, VALIDATION OF THE MEASUREMENT PROCESS, Amer Chem Soc Sympos Ser #63 (1977).
[3] Youden, W.J., Anal Chem **32** [13] (1960) 23A.
[4] Currie, L.A., Pure & Appl Chem, **54** (1982) 715.
[5] Lamparski, L.L., and T.J. Nestrick, Anal Chem, **52** (1980) 2045.
[6] Kirchhoff, W.H., Edit, NBSIR 82-2511 (1982).
[7] Parr, R.M.; H. Houtermans, and K. Schaerf in COMPUTERS IN ACTIVATION ANALYSIS AND GAMMA-RAY SPECTROSCOPY, US Dept of Energy, CONF-780421 (1979) 544. (See also Zagyvai, P.; R.M. Parr and L.G. Nagy, J. Radioanal. Nucl. Chem. **89** 589 (1985).)
[8] Ritter, G.L., and L.A. Currie, ibid, p. 39.
[9] Nalimov, V.V., FACES OF SCIENCE, ISI Press (Philadelphia, 1981).
[10] IAEA Analytical Quality Control Service Program, LAB/243 (1984).
[11] Currie, L.A.; R.W. Gerlach, C.W. Lewis, W.D. Balfour, J.A. Cooper, S.L. Dattner, R.T. DeCesar, G.E. Gordon, S.L. Heisler, P.K. Hopke, J.J. Shah, G.D. Thurston, and H.J. Williamson, Atm Environ **18** (1984) 1517.
[12] Dzubay, T.G.; R.K. Stevens, W.D. Balfour, H.J. Williamson, J.A. Cooper, J.E. Core, R.T. DeCesar, E.R. Crutcher, S.L. Dattner,

B.L. Davis, S.L. Heisler, J.J. Shah, P.K. Hopke, and D.L. Johnson, Atm Environ **18** (1984) 1555.
[13] Novak, J.H., and D.B. Turner, J. Air Polut. Control Assoc., **26** (1976) 570.
[14] Cheng, M-D., and P.K. Hopke, An Intercomparison of Linear Programming Procedures for Aerosol Mass Apportionment, Air Pollution Control Association Paper No. 85-21.8 (1985).
[15] Frank, I.E., and B.R. Kowalski, Statistical Receptor Models Solved by Partial Least Squares, ACS Divn of Environ Chem Sympos Abstracts (Philadelphia, Aug 1984) p. 202.
[16] Henry, R.C., Proc Air Pollut Control Assoc Spec Conf, **SP-48** (1982) 141.
[17] Lowenthal, D.H.; R.C. Hanumara, K.A. Rahn, and L.A. Currie, Estimates and Uncertainties in Chemical Mass Balance Apportionments: Quail Roost II Revisited, prepared for submission to Atmospheric Environment (1985).
[18] Stevens, R.K., and T.G. Pace, Atmos. Environ. **18** (1984) 1499.

DISCUSSION

of the L.A. Currie paper, The Limitations of Models and Measurements as Revealed Through Chemometric Intercomparison

Leon Jay Gleser
Department of Statistics
Purdue University

The construction and use of Simulation Test Data (STD) to help evaluate alternative chemometric methodologies is a highly welcome contribution to the field. Dr. Currie, and the agencies and colleagues whom he credits, are to be congratulated for an approach which has the potential to promote improvements in the art of quantitative chemical analysis.

What follows is a brief discussion of some previous use of standard data sets in statistical research, along with some warnings about the possible pitfalls connected with the use of such approaches. In particular, the parallel that Dr. Currie draws between the use of standard data sets and interlaboratory comparisons using common reference materials cannot be pushed too far. Many interacting factors lead to bias in modeling and analysis of complex data sets; the contributions of these factors would be confounded in typical interlaboratory comparison designs. One factor, scientific judgment, cannot even be identified in standard frequentist reports of statistical data analysis. This suggests that subjective scientific judgments need to be given more explicit mention in reports of statistical analyses, perhaps through the use of the Bayesian approach to inference.

To make standard data sets more closely resemble real-world data, the use of the "bootstrap" is suggested. The "bootstrap" can also help in providing the estimates of statistical precision that Dr. Currie notes were lacking in the two studies conducted to date.

Standard Data Sets in Statistics

Statisticians have long recognized the usefulness of having common data sets on which new methodologies can be tried out, and their relative merits assessed. For example,