NIST Technical Note 2114

Developing Cost Functions for Estimating Solar Photovoltaic System Installed Using Historical Data and OLS Regression

David Webb
Joshua Kneifel

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2114

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

NIST Technical Note 2114

Developing Cost Functions for Estimating Solar Photovoltaic System Installed Using Historical Data and OLS Regression

David Webb
Joshua Kneifel
Office of Applied Economics
Engineering Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2114

November 2020

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

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.

National Institute of Standards and Technology Technical Note 2114
Natl. Inst. Stand. Technol. Tech. Note 2114, 42 pages (November 2020)
CODEN: NTNOEF

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2114

Abstract

Solar photovoltaics (PV) continues to increase in market share. Policy decisions and the nature of solar markets continue to shift; however, it is likely that the price of solar will continue to decrease in the near term. Given the increasing market and more competition in installations, it is beneficial to have a greater understanding in the driving factors in solar PV pricing, as well as models to help perspective buyers and sellers to obtain estimates for the cost of installations. Currently, most estimates rely on a marginal cost that is equivalent to the total cost divided by the system size. This study uses data from EnergySage and the National Renewable Energy Laboratory's Tracking the Sun data set for California, specifically Fresno, San Francisco, Los Angeles, San Diego, and San Jose, to accomplish three goals: to determine if there are significant predictors for solar PV pricing outside of the current method of relying on system size only, to determine what model would make sense for predictive purpose in preparation for the development of a tool to predict the total life cycle cost of solar PV, and to determine if smaller geographical resolutions are warranted when looking at price by location. This paper finds that there are several more significant predictors of Solar PV pricing by including more PV system specifications, such as panel efficiency, inverter type, and system quality. Results also indicate that the installer of the PV system may proxy for the specification variables when it is included in the model. While the installer-based models show significant difference from many of the other models, including the specification-based models, they fail to increase the predictive capability for the EnergySage data, however, show promise for better predictions using the Tracking the Sun data. This difference is driven by the EnergySage data being far more dependent on system size to the point that it can serve reliably as a quote predictor on its own. By breaking the data down to models by city and city-installer groups regional differences can be clearly seem, indicating a more refined geographic approach is necessary for PV price estimation.

Key words

Cost Estimation; Economics; Regression Analysis; Solar Photovoltaics.

Table of Contents

List of Tables

List of Figures

1. Introduction

Solar photovoltaic (PV) system installations for residential homes have expanded significantly since 2010. Analysis from the National Renewable Energy Laboratory (NREL) finds that total installations per year in the United States increased from less than 50 000 in 2010 to over 350 000 in 2016 [1]. The data indicates a dip in 2017, but still over 300 000 new systems were installed in both 2017 and 2018. Given the increasing prevalence of solar PV, economic analysis (both current and projections) of solar PV systems is becoming increasingly important to understand the nature of the market.

A key driver of the growing deployment of residential solar PV systems has been the decrease in the installed cost to a homeowner. The reported national median installed cost of residential solar PV systems has decreased from nearly $10/W in 2008 to ~$3.70/W in 2019 [1]. The average cost has decreased due to reductions in costs for all cost categories (PV panels, inverters, balance of systems (BoS), and "soft costs" such as customer acquisition and margins) as well as economies of scale from larger median array installations (grown from ~4.2 kW_{DC} in 2008 to 6.4 kW_{DC} in 2018) and improved technology such as higher median efficiency panels (grown from 14 % in 2008 to over 18 % in 2018). EnergySage data shows that the downward trend in prices and increasing size of residential solar PV arrays appears to have continued in 2018 and 2019 as the quoted average installed costs of $3.05/W with average system size of 9.6 kW in the second half of 2018 and a further reduction in cost in 2019 thus far at $2.96/W [2]. It's important to keep in mind that quoted prices do not necessarily translate into the installed price since the installed price may be impacted by unexpected costs or delays in the design, installation and permitting process.

The quoted prices have been consistently lower than the reported realized installed costs for a given year by $0.36/W to $0.54/W (9 % to 14 %), which could be driven by numerous factors. We will highlight two here. First, the two prices may be capturing different types of customers and markets. Second, the quoted prices represent potential future system installations that may not be reported for one or two years. When comparing the reported median installed costs to the average quoted cost, the quoted estimates appear to be a relatively good projection for future reported installation costs using a 2-year lag as shown in Figure 1.

Based on technical modeling, NREL has estimated the engineering-based benchmark (technically feasible) price to be $2.70/W [3]. The modeled benchmark installed costs has been consistently below the reported installed cost ($0.73/W to $0.98/W) since 2013. Their benchmark cost has been decreasing at a slower rate year-over-year ($0.14/W in 2018) as the installed prices get closer to the technically feasible cost estimates. Assuming a 3-year lag on the benchmark to align it with the installed and quoted costs can be used for a projection of future average installed costs (Figure 1).

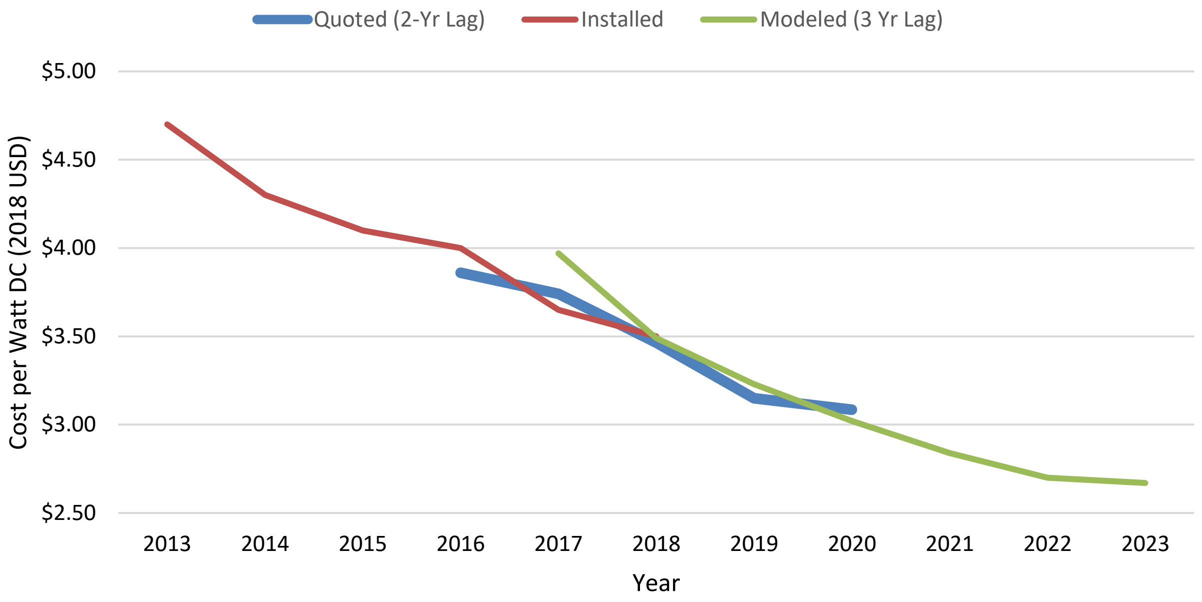

Figure 1. Installed Cost versus Quoted Cost (2-Year Lag) versus Modeled (3-year Lag)

Although this national trend is important, the decision to install a solar PV system is specific to factors related to a homeowner's location. Barbose et al. [1] shows that the median installed cost across 20 states in 2018 ranges from $2.80/W to $4.40/W. Similarly, EnergySage [2] shows the average quoted price for 36 states ranging from $2.66/W to $3.29/W through 2019. There is a potential for even greater market variation across administrative and jurisdictional lines (county, city, or neighborhood level). These differences are a result of numerous factors, including customer demand/awareness, market development stage, state and local labor rules, laws, and regulations, and other regional effects.

To date, cost data has typically been reported on an average cost per watt basis. This approach makes sense when most of the costs are associated with each installed watt (solar panels and inverters). However, as these costs have become smaller, there is potential for costs not directly associated with the size of the system (fixed costs, costs associated with the complexity of the system, differences in system quality) to account for a greater share of overall costs. For example, the median reported installed price for a system with 18 % to 19 % efficiency panels is $3.60/W versus $4.00/W for 20 % to 21 % efficiency panels [1]. Fixed costs (e.g. customer acquisition costs, permitting and commissioning) may vary based on the state or county system approval processes and the awareness of customers. Markets that are well developed with multiple installers realize lower margins, and therefore lower installed costs to homeowners [4].

Differences in the market may also play a role. Barbose et al. [1] accounts for this at the state level, however finer gradations may be more appropriate as county and local level ordinances and permitting may alter the costs of installation. Other potential factors affecting price include the specific installer and the specifications of the system itself. Tracking the Sun does not examine the former; however, it does attempt to account for the latter by using a quality variable. This variable is determined through a combination of factors including system efficiency, warranty, and reliability.

This study uses two data sets to examine the possibility of fixed cost impacts on pricing, more refined localities, installer effects, and specifications. The first is the publicly available Tracking the Sun (TTS) data set used by NREL. This provides all of the data used in the development of Barbose et al. [1]. The second is a privately-owned collected data set from EnergySage, an online marketplace supported by the U.S. Department of Energy where homeowners and businesses can comparison shop custom solar quotes from pre-screened solar contractors.

Given the large data sets involved and the numerous variables in each an ordinary least squares (OLS) regression is used as an initial probe of the data set. There is a high degree of linearity in the data sets, though accompanied by extensive heteroskedasticity, making OLS useful as an initial foray into the data. The goal is to determine the key regressors in the OLS context and use that to inform future, more complex models, and to determine areas where expanded datasets may be appropriate. This study builds on previous work focusing on the DC-Maryland-Virginia region which can be found in Webb et al. [5].

2. Literature Review

Several organizations provide installed cost data for residential solar PV systems, most notably the Lawrence Berkeley National Laboratory (LBL), the National Renewable Energy Laboratory (NREL), and EnergySage. NREL provides the annual Tracking the Sun report [1] and have published numerous reports and journal articles evaluating solar PV market structure (O'Shaughnessy (6), O'Shaughnessy (7)).   NREL reports contain trends analysis in technology installation including recent historical data (1 to 2 years old) and modeled engineering-based (technically feasible) cost estimates. EnergySage provides bi-annual summaries of installer quotes provided in its online customer platform. The key specifics of included data are found in the Methodology section, but they include varying technology options, locations of the system, size of the system, among various other energy, engineering, location, and financial information. Quotes are more representative of current and near-term future installed costs because they are estimates for systems not yet installed. Using this data provides a reasonable expected installed cost for the next year, providing current or forward-looking analysis as opposed to backward looking (historical) analysis.

These resources are insightful into the general trends of the installed cost of residential markets for solar PV in the United States but are generalized over large markets in most cases and focus only on installation costs. The monetary benefits of solar PV are dispersed over the life of the system and some costs do not accrue immediately (maintenance, replacement, grid access fees and tariffs). Economic analysis can properly account for these future costs and many prior studies have evaluated the net present value (NPV) and internal rate of return (IRR) of residential solar PV.

An older case study in Denmark found that investments in energy efficiency were more effective than in renewable technologies [8]. Solar PV with a heat pump was cost-optimal for a Net-Zero structure in a dense city area while solar PV with district heating is the highest lifecycle cost (LCC) due to high operation and maintenance costs. In terms of energy efficiency, the best performing system was a solar PV system coupled with a solar thermal system and solar heat pump, although it was not optimal in terms of LCC. Another study in Canada found that solar PV could not achieve payback in 60 years unless the initial price of electricity increased by greater than 5 % per year using a 4 % discount rate [9]. This increases to 78 years with a higher discount rate equal to the inflation rate. A study with a focus on Singapore reached similar conclusions, finding residential PV to have a higher LCC than utilizing grid-based electricity [10].

More recent studies have found solar PV to be more economically viable. Swift (11) examined the economics of solar PV by looking at locations across the United States, including specific incentives, electricity rates, and solar insolation. The IRR ranged from 31.6 % in Honolulu to 8.3 % in Minneapolis. By varying the installed cost of solar PV, the authors also estimated the required installed cost to make solar PV economically attractive based on IRR. Parity with grid produced electricity with and without incentives was found to be location specific. A study published in 2015 found that PV was an attractive investment in many countries even in the absence of incentives [12], once again showing highly location specific variability. Farias-Rocha, Hassan, Malimata, Sánchez-Cubedo and Rojas-Solórzano (13) examined the economic feasibility of solar PV in the Philippines by focusing on the minimum feed-in tariff, the viability of net metering, and any additional support mechanisms that would be useful for supporting solar PV. The authors found that a 100 kW feed-in tariff would be profitable for a solar investor if the tariff does not drop below 4.20 PHP/kWh. A 1.89 kW system was found to be financially attractive using net metering alone. A recent Canadian study examining urban deployment of rooftop solar PV found 96 % of identified suitable rooftops would be profitable using NPV [14]. Recent studies in India have found solar PV to be financially viable for residential systems [15] and rural areas [16] while a study in Spain found utilizing grid electricity and natural gas for heating to be more economical than solar PV coupled with solar thermal and a micro-CHP system [17]. A more recent study for the United States by Lee, Hong, Koo and Kim (18) found that 18 states realized a payback period to at least break even while the other 32 states not being able to reach a breakeven point. Depending on the state and incentive, the payback period for those that at least broke even ranges from as high as 25 years (Nevada and Wisconsin) to as few as 5 years (Hawaii). Maryland and Washington DC had payback periods of 18 and 10 years, respectively. The focus of this study is California, which did not reach breakeven over the lifetime of the solar PV system. These differing results indicate both the improving economics of residential solar PV systems and the impact of state and regional differences when examining the LCC of solar PV systems.

Several studies also examine the impact of various incentives on the economics of solar PV. A study for the European Union examined the impact of various incentives, such as feed-in tariffs, net metering, capital subsidies, grants and rebates, and green tags [19]. The study examined multiple countries for both wind and solar PV, finding that depending on what incentives were available and how they were implemented, incentives can vary from beneficial to inconvenient for renewable energy sources. A partial rework by the authors expanded the number of countries considered and focused solely on feed-in tariffs finding the same basic results [20]. This finding is echoed in Dusonchet and Telaretti (21). Sow, Mehrtash, Rousse and Haillot (22) found that, for Canada, incentives allowed projects to remain feasible (based on 2016 data) with the only exception being projects in Montreal.

United States based studies also include the examination of Solar Renewable Energy Credits (SRECs). Burns and Kang (23) examined the early state of many SREC markets, finding them to be potentially strong, though the uncertainty associated with them proved to be a major drawback. Specifically, the SREC market had a higher present value than any other incentive examined (ITC, net metering, state tax credits), but the fluctuation in prices meant any benefit was highly uncertain. At the time of the study (2012) these benefits had a variable effect based on energy price, with less incentive required when net metering was available, while solar PV in Ohio was still not economically competitive due to the state's lower energy prices. An analysis examining uncertainty in the cost-effectiveness of residential solar PV found that incentives that reduce the uncertainty in solar PV returns were generally the most effective [24]. The study, focused on Massachusetts, found uncertainties that lead to delays in investment timing and the discounted benefits of solar PV needed to exceed investment cost by 60 % to trigger investment. A study focusing on the United States as a whole found that the impact of incentives lead to a highly variable profitability index by state [18].

Work done in Webb et al. [5] found that the inclusion of a regression constant to account for fixed costs produced statistically significant differences in the mean of the regression for smaller and larger systems. Specifically, systems much smaller than the mean sized system tended to be underestimated in terms of cost when using the marginal only model and systems much larger than the mean sized system tended to be overestimated. The constant was found to be significant in the regression and given the large amount of data near the origin indicated that the fixed cost component warrants inclusion. Webb et al. [5] also applied the regression results to a lifecycle cost analysis, finding minor differences in total LCC when examining different counties in the Washington D.C.-Maryland-Virginia region of the United States.

This study has three goals: to determine if there are significant predictors for solar PV pricing outside of the current method of relying on system size only, to determine what model would make sense for predictive purpose in preparation for the development of a tool to predict the total life cycle cost of solar PV, and to determine if smaller geographical resolutions are warranted when looking at price by location.

3. Data and Methodology

3.1. EnergySage Dataset

The analysis uses a unique dataset provided by EnergySage [25]. EnergySage aggregates quotes for solar installations from multiple solar PV installers provided to homeowners on its online platform for January 2013 to present, although the data for this analysis is limited to California 2018. Versions of this dataset have been used before [26], but the current analysis is fundamentally different because it focuses on sub-state analysis, the value of models for predictive purposes, and looks at more variables in the regression.

The dataset includes several variables (variable name used in this paper in italics) for each quote, the most pertinent to the current analysis being:

- Quote Date (*Year*)
- System Size in Watts (*Size*) – Direct Current in Watts (W_{DC})

- Quote for Purchase Price (*Quote*) in USD[1]
- System Quality in Six Qualitative Tiers: economy, economy plus, standard, standard plus, premium, premium plus (*Tier*)
- City (*City*)
- ZIP Code (*ZIP*)
- Installer (*Inst*)
- Inverter Type (*Inv*)

The data was anonymized in terms of physical address of the property and the name of solar installer for the purposes of this report.

There were issues with the data due to the voluntary nature of the input.

1. System Quality (*Tier*) is not consistently reported for all years and occasionally within tiers
2. Tiers do not always have a sufficient number of data points to allow analysis
3. Some quotes do not contain a quote price

To address the first issue a separate category for any non-tier list system is created and labeled Tier 0. This leads to the possibility of a mixture of systems in the Tier 0 category, and therefore the Tier 0 system quotes are excluded from any analysis that includes the tier variables. The second issue is resolved by aggregating the provided tiers (non-Tier 0 labeled quotes) into a three-tier classification of economy with economy plus (labeled *Economy* from this point on), standard with standard plus (labeled *Standard* from this point on), and premium with premium plus (labeled *Premium* from this point on). Given the prevalence of standard and premium systems, there were not enough economy system quotes to include in the analysis, and therefore, are excluded.[2] Issue three required dropping the no value quotes from the analysis as there was no way to determine the true value of the quoted system.

Data was provided for all EnergySage quotes for California in 2018. The analysis focuses on rooftop residential solar PV and excludes non-residential systems or those whose mounting system was not “penetrating rooftop” from the analysis.

Three types of inverters appear in the data set after filtering: *Micro*, *String*, and *Optimizer*. Most systems quoted in 2018 have either a microinverter or optimizer as part of a string inverter. Additionally, systems with optimizers and microinverters have similar overall installed costs [1]. Therefore, this restriction should be a reasonable representation of the market systems and costs. A further filter was applied to remove those systems over 30 000 W_{DC} to account for overly influential points in sparse data regions as well as erroneous data entries relative to the defined filters. 30 000 W_{DC} is also the largest a system can be in California before a special exemption is required to have the system treated as a residential system.

Based on the literature, there are several variables that have a clear expected impact on installed costs. Larger system size and higher quality systems are expected to increase installed costs. Systems with string inverters without optimizers are expected to be cheaper

[1] Quotes are used because reported installed prices are not available; a quote does not always end in a purchase.

[2] Typical solar PV panel efficiencies and production quality have been consistently increasing year-over-year and the trend is expected to continue, leading to minimal installations of “economy” or low efficiency panels.

than systems with microinverters or optimizers. However, there is less clarity on whether these variables will influence the marginal cost, fixed cost, or both.

3.2. Tracking the Sun Dataset

Tracking the Sun is a yearly publication produced by NREL that examines trends in solar PV pricing. It leverages installed prices across participating agencies throughout the United States representing a roughly 80 % of the domestic solar PV market [1]. The TTS data set is publicly available and therefore locator information is limited to state, city, and ZIP code. Furthermore, since the data set is an aggregation of multiple state and local entity reports, some based on self-assessment by system owners, certain fields are not consistently reported.

Key variables (and expected impacts) in the data set remain principally the same as for the EnergySage data, with a few alterations. The price reported in the TTS data is the installed price and not a quote, thus TTS regressions use the *Price* variable. Inverter type is not reported directly so it must be synthesized using other variables related to the inverters. There is no system quality variable in the TTS data and no reference to the formula used to generate a quality metric comparable to EnergySage, however the module efficiency (*Eff*) is available. *Eff* does not capture all the characteristics captured by the quality variable, which combines multiple factors (efficiency, warranty, and performance) into a single qualitative metric. Therefore, *Eff* may influence installed costs in a different manner than quality influences the quoted costs. As with EnergySage, installer name is anonymized in this analysis, though the public nature of the data set makes it unnecessary. Due to a lack of overlap in installers between the EnergySage and Tracking the Sun datasets as well as the large number of installers represented in each, there is no way to infer the installers represented in EnergySage from the TTS data.

The TTS data was also filtered to ensure both data sets were examining the same system types. Systems installed in years other than 2018, ground mounted systems, systems with battery backups, non-residential systems, tracking systems, systems over 30 kW_{DC}, and systems with the *appraised value flag* were all filtered. The last of these filters is done at the express recommendation of the guidance on using the TTS data. Furthermore, module types that had too few instances in the data to provide statistical results were also filtered.

Attempts to link the EnergySage systems with the TTS systems datasets were unsuccessful due to the lag between 2018 quotes showing up in the installation data for TTS, assuming those quotes show up at all. As such a comparison of the two is infeasible given the currently available data. If more data were available it may be possible to analyze the two data sets together and model how quoted prices translate to installed prices, the rate at which quotes are accepted, and whether there are any systematic differences between prices obtained through a clearing house versus those that from direct sales.

3.3. Statistical Analysis

The study focuses on only five locations in California: San Jose, San Diego, San Francisco, Los Angeles, and Fresno. These cities were chosen because they had enough installations to provide statistically significant results for every model developed in this report. Indicator variables are used where appropriate to analyze differences between groups.

The analysis relies on a series of OLS regressions with robust standard errors to assess the impacts of the variables in each model. Each model is then compared along multiple criteria.

The first is significance of predictors, or in the case of indicator variables, the significance of differences in predictors for indicator groups. Predictor significance informs whether the variable in the model has some statistical relationship with the predicted variable. Second is the adjusted R^2 of the model. The adjusted R^2 measures how much of the variation in the data is explained by the model and is an important measure if interested in the predictive power of the model. However, it is generally not a useful indicator of a model's appropriateness on its own and needs to be supplemented with other tests, for instance cross validation, prediction intervals, or comparisons of the mean squared prediction error (MSPE) between models. Lastly, the prediction and confidence intervals of the model estimates are developed and compared. The former provides evidence of whether it is possible to statistically say that a prediction came from one model with a certain level of significance, while the latter evinces whether the mean predictions from two models differ statistically. Information criterion are also used in selecting between models.

Due to the use of robust standard errors to account for heteroskedasticity, the typical formulas for hypothesis testing do not work. Outside of the significance of predictors, which can be determined using the Huber-White Sandwich estimator, nonparametric bootstrapping is utilized for differences in adjusted R^2 values between models and confidence intervals on the line as well as other regression statistics that require adjustment due to the use of robust standard errors, while quantile regression is utilized for prediction intervals.

4. Analysis

4.1. EnergySage Regressions

The primary driver for a quoted system cost remains the size in W_{DC} of the system in question. As such, most estimates of solar PV price use only the average based on system size when developing estimates (essentially the mean total cost per watt). Conceptually, one could argue that if there is no system then there is no cost of installation and no need to add a fixed cost, however the model is predicated on a system being installed, as such the fixed cost of installation should be evidenced in any model. For the purposes of the initial inspection of the EnergySage and TTS data, this paper assumes linearity through the entire data region through using OLS, although it is possible that the fixed cost may induce some non-linearity near the origin. Webb et al. [5] presents a justification for inclusion of the regression constant on the basis that certain costs are not on a per watt basis, however there are also statistical reasons to include it. In situations where there isn't enough data near the origin to train the model in that region, enforcing no constant can bias the model by assuming a set value where the data cannot statistically justify it.

Bearing the regression constant in mind, the first OLS model assumes the most simplistic form, see Equation 1. At this point the "Economy" tier is dropped from the analysis, as it does not have sufficient observations to maintain significance through all regressions, leaving "Standard" and Premium" tiers. This is a result of panel efficiencies increasing rapidly, as observed in Barbose et al. [1].

$$Quote = \beta_1 * Size + \beta_0 + \epsilon_0 \tag{1}$$

Where β_1 is the regression coefficient on system size, β_0 is the regression constant, and ϵ_0 is the error term for the model. For simplicity, all future models use β to represent coefficients, though they are not equal. In this case the error term does not meet the requirements of the basic OLS model as the data has a high degree of heteroskedasticity, as evident in Figure 2. Note that all regressions performed on the EnergySage data are based on the same set of 9357 data points.

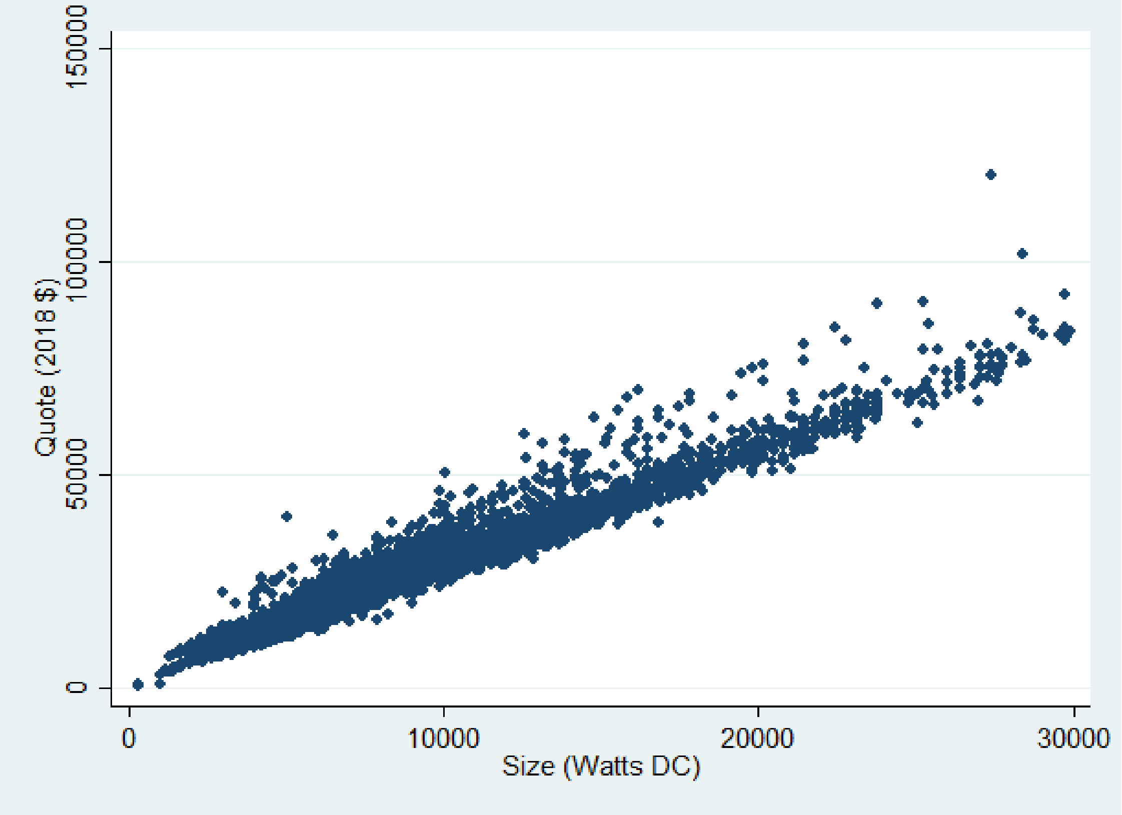

Figure 2. Plot of Quoted Price against System Size for the Filtered EnergySage Data

The results of the regression using Equation 1 are presented in Table 1. The adjusted R^2 is 0.9475 with a marginal price[3] of \$2.79/W_{DC} and a fixed cost of approximately \$1500. This represents a high degree of linearity indicating quotes may be based on fairly simple cost models. From this basic model two different models are developed, one based on system specifications and another based on location and installer.

Table 1. Equation 1 Regression Results

	Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
Size	2.788	0.006	500.700	0.000	2.777	2.799
Constant	1523.913	47.791	31.890	0.000	1430.235	1617.591

4.1.1. Advanced System Specification Models

The first iteration of the solar PV system specification model adds the quality variable interacted with size as an indicator, as seen in Equation 2. Quality is directly related to the

[3] Marginal cost refers to cost per additional Watt. Solar PV panels typically come in non-divisible units (i.e. a 320 W panel) making actual panels more akin to lump sum payments. There is no standard panel size though, so the marginal cost is used.

solar PV panel (e.g., efficiency) and, therefore, is expected to primarily impact the marginal cost. However, there are other factors expected to be captured in quality that could impact the fixed costs (e.g., warranties).

$$Quote = \beta_1 * Size + \beta_2 * Size\#i.Tier + \beta_3 * i.Tier + \beta_0 + \epsilon_0 \qquad (2)$$

β_0 is the regression constant, all other β values are coefficients, and # represents the interaction between two variables. $i.variable\ name$ means the variable is an indicator variable.

Table 2 contains the regression results. Adding quality creates a new significant predictor to the model but has little impact on the adjusted R^2 at 0.9480. Looking at the mean squared prediction error (MSPE) using an 80/20 training to test split reveals no statistical difference between the two models (Equation 1 MSPE is 7 106 548, Equation 2 MSPE is 6 953 777)[4]. The "Premium" panels add $251 in fixed costs and $0.06/W in marginal costs over "Standard" panels. The data reveals a significant relationship with quality but no effect on prediction. For the goal of prediction, the additional increase in predictive power is not justified by the additional model complexity.

Table 2. Equation 2 Regression Results

	Robust Coef.	Robust Std. Err.	t	P>t	[95% Conf.	Interval]
Size	2.790	0.0105	266.12	<0.001	2.770	2.811
Tier						
Standard	-251.037	108.162	-2.32	0.020	-463.05	-39.025
Tier#Size						
Standard	-0.0615	0.0173	-3.55	<0.001	-0.0955	-0.0276
Constant	1612.984	71.975	22.41	<0.001	1471.903	1754.065

The next model excludes the quality variable and includes the inverter type as shown in Equation 3. The inverter type (string versus microinverter) is expected to impact the costs of a solar PV system, potentially through both marginal cost because the size of the inverter is directly correlated with the size of the system, and fixed costs because different inverter approaches may require different hardware and labor costs not associated with the size of the system.

$$Quote = \beta_1 * Size + \beta_2 * Size\#i.Inv + \beta_3 * i.Inv + \beta_0 + \epsilon_0 \qquad (3)$$

Table 3 contains the results of the regression. Similar to the quality variable, the inverter type is significant, with microinverters being less expensive by $1060 in fixed cost relative to string inverters and more expensive in marginal cost by $0.06/W. In terms of explained variance or prediction error, including inverter variables does not add predictive power of the model (adjusted R^2 is 0.9479 and MSPE is 7 046 423).

[4] All comparisons of adjusted R^2 and MSPE are done using a non-parametric bootstrap with 100 resamplings

Table 3. Equation 3 Regression Results

	Robust Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
Size	2.840	0.0207	137.22	<0.001	2.800	2.881
Inv						
String	1060.241	160.938	6.59	<0.001	744.781	1375.701
Inv#Size						
String	-0.0640	0.0231	-2.77	0.006	-0.109266	-0.0187
Constant	668.922	146.366	4.57	<0.001	382.026	955.819

The final specification that is possible to regress on in the EnergySage data is the use of a DC optimizer in conjunction with a string inverter. This regression is similar to Equation 3; however, the string inverter systems must be further disaggregated.

Table 4 presents the regression results. The string inverter without a DC optimizer (*String No Opt*) shows no significant difference from microinverters in both marginal and fixed cost, which could be driven by a lack of data points for that system type. There is a significant difference from a string inverter with a DC optimizer relative to the microinverter ($1056 high fixed costs and $0.06/W lower marginal costs). This result is logical because solar PV panels with microinverters built into the panel tend to be more expensive than panels without microinverters, but do not require the hardware and labor to install string inverters. The inverter variables may also be capturing the impact of increased efficiency of the solar PV panels because microinverters are typically included in high efficiency panels. Explained variance (EV) and prediction error (PE), however, is not significantly approved (adjusted R^2 = 0.9479, MSPE = 7050633).

Table 4. Equation 3 Regression Results with DC Optimizer

	Robust Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
Size	2.840	0.021	137.210	0.000	2.800	2.881
InvOpt						
String No Opt	1740.260	1113.664	1.560	0.118	-442.670	3923.191
String Opt	1055.893	161.005	6.560	<0.001	740.302	1371.483
InvOpt##Size						
String No Opt	-0.100	0.187	-0.540	0.590	-0.466	0.265
String Opt	-0.064	0.023	-2.760	0.006	-0.109	-0.018
Constant	668.922	146.376	4.570	<0.001	382.006	955.839

For predictive purposes, the system size is sufficient for predictive purposes when using the EnergySage data, however there are additional model specifications available using the available variables. The final regression involves finding a model that includes the most significant predictors. After examining multiple model specifications, Equation 4 was developed and includes size, tier, tier interacted with size, inverter, and inverter interacted with size. The inverter variable in this case includes the DC optimizer option.

$$Quote = \beta_1 * Size + \beta_2 * Size\#i.Tier + \beta_3 * i.Tier + \beta_4 * Size\#i.Inv + \beta_5 * i.Inv + \beta_0 + \epsilon_0 \tag{4}$$

The results of this regression are presented in Table 5. The size, tier, and inverter type are all statistically significant, although the Tier-Size interaction is only marginally statistically significant at the 90 % confidence level. As before, the EV and PE are not significantly improved (adjusted $R^2 = 0.9485$, MSPE = 6 900 347). Otherwise, the significance trends from the previous models are preserved. A model involving a triple interaction between the three predictors was examined, but most predictors lost significance.

Table 5. Equation 4 Regression Results

	Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
Size	2.784	0.021	131.840	0.000	2.743	2.826
Tier						
Standard	-377.268	179.312	-2.100	0.035	-728.808	-25.729
Tier#Size						
Standard	-0.051	0.028	-1.810	0.070	-0.105	0.004
Inv#Size						
String No Opt	-0.376	0.174	-2.160	0.031	-0.717	-0.035
String Opt	-0.006	0.029	-0.210	0.833	-0.062	0.050
InvOpt						
String No Opt	3409.467	1131.085	3.010	0.003	1191.984	5626.950
String Opt	650.080	182.706	3.560	0.000	291.888	1008.273
Constant	1162.598	150.517	7.720	0.000	867.511	1457.686

The benefit of this model is that is allows for comparison of different system configurations. Table 6 shows the estimated fixed cost and marginal cost based on the different configuration options. Let's compare the following: premium panels with microinverters, standard panels with string inverter and optimizers, and standard panel with sting inverter. The fixed costs for these three configurations are $4195, $1435, and $1163, respectively, while the marginal costs are $2.36/W, $2.73, and $2.78/W. The premium system with microinverters has statistically significant lower fixed costs and higher marginal costs.

Table 6. Estimated Fixed Cost and Marginal Cost by System Specification

Fixed Cost	**Standard**	**Premium**
String	4195	4572
Opt	1435	1813
Micro	785	1163

Marg Cost	**Standard**	**Premium**
String	2.357	2.408
Opt	2.727	2.778
Micro	2.733	2.784

Assuming a 10.0 kW system, the installed costs are estimated at $27 765, $28 765, and $29 003, respectively. The $1238 difference in installed costs would not be captured in the Size only model as all these systems would have the same predicted value.

4.1.2. Installer Models

Instead of focusing on the specifications of the system being installed, the installer model focuses on two variables in the EnergySage data set, City and Installer. The installer is expected to capture some of the same variation identified by quality tier and inverter characteristics as well as installer specific cost variation. The city is expected to capture market cost and competition differences. One city may have a more developed market with more competition and more informed consumers, and lower costs driven by installers being further out the learning curve. Also, some cities may have more stringent permitting and commissioning processes that increase installation costs. Three regressions are performed, the first focusing on just the Installer variable, the next on just the City variable, and the final regression examining a City-Installer group variable. All regressions include system size. These regressions are given in Equations 5 through 7.

$$Quote = \beta_1 * Size + \beta_2 * Size\#i.Inst + \beta_3 * i.Inst + \beta_0 + \epsilon_0 \tag{5}$$

$$Quote = \beta_1 * Size + \beta_2 * Size\#i.City + \beta_3 * i.City + \beta_0 + \epsilon_0 \tag{6}$$

$$Quote = \beta_1 * Size + \beta_2 * Size\#i.CityInst + \beta_3 * i.CityInst + \beta_0 + \epsilon_0 \tag{7}$$

Given the large number of installers, many smaller installers have too few installations to produce statistically significant results. Therefore, this analysis focused on the top 10 installers (representing 65 % of all quotes) to keep comparisons interpretable.

Table 7 presents these results. The installer model does show a significant improvement in adjusted R^2 (0.9859) and MSPE (5 363 717) and significant predictors for the selected top 10 installers. Fixed costs vary across the installers by $2346 ($1465 lower to $881 higher) relative to the base installer and marginal costs varying by $1.22/W_{DC}$ relative to the base installer. However, when examining the prediction interval compared to the Size only model, the prediction intervals (estimated using quantile regression) overlap for over 99 % of the data points. Thus, if using the model for predictive purposes there would be no statistically significant way to claim that a prediction from the installer model could not have also come from the Size only model.

Table 7. Equation 5 Regression Results

	Coef.	Robust Std. Err.	T	P>t	[95% Conf. Interval]	
Size	2.452	0.013	194.690	<0.001	2.427	2.476
Inst						
2	-1302.196	80.527	-16.170	<0.001	-1460.048	-1144.344
3	-309.000	102.869	-3.000	0.003	-510.645	-107.354
4	880.937	257.051	3.430	0.001	377.060	1384.814
5	-549.344	104.107	-5.280	<0.001	-753.418	-345.270
6	-1102.930	163.586	-6.740	<0.001	-1423.596	-782.265
7	-1270.441	136.809	-9.290	<0.001	-1538.618	-1002.264
8	-1465.251	105.029	-13.950	<0.001	-1671.133	-1259.370
9	-732.335	119.591	-6.120	<0.001	-966.761	-497.910
10	-195.809	260.075	-0.750	0.452	-705.616	313.998
Inst#Size						
2	0.244	0.013	18.070	<0.001	0.218	0.270
3	0.306	0.016	19.270	<0.001	0.275	0.337
4	0.219	0.030	7.260	<0.001	0.160	0.279
5	0.311	0.017	17.840	<0.001	0.277	0.346
6	0.452	0.030	15.230	<0.001	0.394	0.510
7	0.376	0.021	18.060	<0.001	0.335	0.417
8	0.402	0.016	24.610	<0.001	0.370	0.434
9	1.216	0.021	57.520	<0.001	1.175	1.258
10	0.188	0.030	6.220	<0.001	0.129	0.247
Constant	1811.129	72.221	25.080	<0.001	1669.560	1952.699

The City model is presented in Table 8 and, as with the installer model, only looks at the top 10 installers to allow comparison with other installer models.

Table 8. Equation 6 Regression Results

	Coef.	Robust Std. Err.	T	P>t	[95% Conf. Interval]	
Size	3.049	0.072	42.380	<0.001	2.908	3.190
City						
Los Angeles	-312.693	558.815	-0.560	0.576	-1408.077	782.690
San Diego	206.677	576.463	0.360	0.720	-923.300	1336.654
San Francisco	576.953	970.231	0.590	0.552	-1324.884	2478.790
San Jose	-34.950	585.461	-0.060	0.952	-1182.566	1112.666
City#Size						
Los Angeles	-0.269	0.072	-3.710	<0.001	-0.411	-0.127
San Diego	-0.290	0.078	-3.750	<0.001	-0.442	-0.138
San Francisco	0.180	0.158	1.140	0.256	-0.130	0.489
San Jose	0.025	0.079	0.310	0.753	-0.131	0.181
Constant	957.548	554.076	1.730	0.084	-128.547	2043.642

The City interaction does have a statistically significant lower marginal cost for Los Angeles (-$0.27/W) and San Diego (-$0.29/W) relative to Fresno (base city) while there is no statistically significant impact on fixed cost for any city. The lower marginal costs may be due to numerous factors, namely the inclusion of different installers for each city. Based on this, a city only model can be specified as in Equation 8. The EV and PE are not significantly improved in this case (adjusted $R^2 = 0.9580$, MSPE = 5 814 342).

Combining the city and installers into a single variable yields Equation 8 and the regression results in Table 9[5].

$$Quote = \beta_1 * Size + \beta_2 * Size\#i.City + \beta_0 + \epsilon_0 \qquad (8)$$

For those variables that are statistically significant (95% CI), the fixed costs vary by $4645 and the marginal costs vary by $1.02/W. While the installer-city model does have a statistically significant improvement in EV and PE in relation to the city only model (adjusted $R^2 = 0.9901$, MSPE = 3 124 373) most of the coefficients are not significantly different from the base city-installer group. For that reason, the city-installer model is an inefficient model for prediction as it is impossible to attribute the increase in predictive power to genuine trends in the coefficients and noise in the data. It should be noted that it appears that the installer may proxy for city, as installers in the current data set remain highly localized to a single city. This also is why city is not treated as an isolated independent variable.

[5] An examination of installers across cities reveals installers mostly stick to markets, making trends in installers across cities difficult to model using the current data set.

Table 9. Equation 7 Regression Results

	Coef.	Robust Std. Err.	T	P>t	[95% Conf. Interval]	
Size	2.825	0.072	39.260	<0.001	2.684	2.966
CityInst						
Fresno 2	1063.126	676.231	1.570	0.116	-262.440	2388.692
Fresno 3	-455.568	688.232	-0.660	0.508	-1804.659	893.524
Fresno 4	-1305.039	671.779	-1.940	0.052	-2621.879	11.801
Fresno 5	-670.635	603.708	-1.110	0.267	-1854.041	512.771
Fresno 6	534.669	656.090	0.810	0.415	-751.417	1820.754
Los Angeles 1	-820.220	578.374	-1.420	0.156	-1953.965	313.525
Los Angeles 2	-704.554	583.913	-1.210	0.228	-1849.156	440.048
Los Angeles 3	-548.900	632.441	-0.870	0.385	-1788.628	690.829
Los Angeles 4	-177.478	601.722	-0.290	0.768	-1356.991	1002.034
Los Angeles 5	-1089.277	613.822	-1.770	0.076	-2292.507	113.953
Los Angeles 6	-1122.307	605.480	-1.850	0.064	-2309.185	64.571
Los Angeles 7	-1369.615	580.865	-2.360	0.018	-2508.242	-230.988
Los Angeles 8	225.991	1159.177	0.190	0.845	-2046.261	2498.242
Los Angeles 9	-219.361	749.934	-0.290	0.770	-1689.402	1250.680
San Diego 1	469.359	580.895	0.810	0.419	-669.328	1608.046
San Diego 2	-776.189	579.290	-1.340	0.180	-1911.730	359.352
San Diego 3	-241.954	596.833	-0.410	0.685	-1411.883	927.976
San Diego 4	2120.894	789.956	2.680	0.007	572.399	3669.388
San Diego 5	-195.160	586.768	-0.330	0.739	-1345.359	955.039
San Diego 6	-716.555	611.518	-1.170	0.241	-1915.268	482.159
San Diego 7	-193.133	595.371	-0.320	0.746	-1360.195	973.929
San Diego 8	-1200.289	581.106	-2.070	0.039	-2339.388	-61.189
San Diego 9	443.965	585.248	0.760	0.448	-703.254	1591.185
San Diego 10	1280.529	657.400	1.950	0.051	-8.126	2569.183
San Francisco 1	-323.729	691.277	-0.470	0.640	-1678.790	1031.332
San Francisco 2	2197.943	1308.913	1.680	0.093	-367.824	4763.710
San Francisco 3	35.488	725.285	0.050	0.961	-1386.235	1457.211
San Francisco 4	-2523.912	617.543	-4.090	0.000	-3734.438	-1313.387
San Francisco 5	1257.294	1297.448	0.970	0.333	-1286.000	3800.588
San Jose 1	529.593	600.579	0.880	0.378	-647.679	1706.865
San Jose 2	494.378	1004.321	0.490	0.623	-1474.321	2463.076
San Jose 3	-43.618	609.091	-0.070	0.943	-1237.575	1150.338
San Jose 4	-1125.979	587.652	-1.920	0.055	-2277.910	25.953
San Jose 5	-1142.989	597.865	-1.910	0.056	-2314.940	28.962

San Jose 6	-480.489	581.991	-0.830	0.409	-1621.323	660.345
San Jose 7	78.679	704.841	0.110	0.911	-1302.971	1460.328
CityInst#Size						
Fresno 2	-0.023	0.080	-0.290	0.775	-0.180	0.134
Fresno 3	-0.052	0.083	-0.630	0.532	-0.214	0.111
Fresno 4	0.132	0.085	1.550	0.121	-0.035	0.299
Fresno 5	0.910	0.075	12.090	<0.001	0.762	1.057
Fresno 6	-0.270	0.079	-3.430	0.001	-0.425	-0.116
Los Angeles 1	-0.129	0.072	-1.780	0.074	-0.270	0.013
Los Angeles 2	-0.034	0.073	-0.470	0.641	-0.176	0.109
Los Angeles 3	-0.077	0.078	-0.990	0.324	-0.229	0.076
Los Angeles 4	-0.062	0.075	-0.820	0.410	-0.208	0.085
Los Angeles 5	-0.043	0.078	-0.550	0.582	-0.195	0.110
Los Angeles 6	0.036	0.075	0.470	0.635	-0.112	0.183
Los Angeles 7	0.032	0.072	0.440	0.657	-0.110	0.174
Los Angeles 8	0.349	0.160	2.190	0.029	0.036	0.662
Los Angeles 9	-0.111	0.086	-1.290	0.197	-0.281	0.058
San Diego 1	-0.373	0.073	-5.110	0.000	-0.516	-0.230
San Diego 2	-0.143	0.073	-1.960	0.050	-0.285	0.000
San Diego 3	-0.084	0.078	-1.080	0.279	-0.237	0.068
San Diego 4	-0.258	0.091	-2.850	0.004	-0.436	-0.081
San Diego 5	-0.054	0.075	-0.710	0.475	-0.202	0.094
San Diego 6	-0.109	0.083	-1.320	0.188	-0.272	0.053
San Diego 7	-0.096	0.077	-1.250	0.211	-0.246	0.054
San Diego 8	-0.024	0.073	-0.330	0.742	-0.167	0.119
San Diego 9	0.643	0.074	8.710	0.000	0.499	0.788
San Diego 10	-0.329	0.081	-4.080	0.000	-0.487	-0.171
San Francisco 1	0.157	0.094	1.660	0.096	-0.028	0.341
San Francisco 2	0.202	0.131	1.540	0.124	-0.055	0.459
San Francisco 3	-0.067	0.098	-0.680	0.497	-0.259	0.126
San Francisco 4	1.471	0.080	18.420	<0.001	1.315	1.628
San Francisco 5	-0.226	0.168	-1.350	0.178	-0.556	0.103
San Jose 1	0.045	0.076	0.590	0.557	-0.105	0.195
San Jose 2	0.098	0.149	0.660	0.510	-0.194	0.390
San Jose 3	-0.040	0.080	-0.490	0.621	-0.197	0.118
San Jose 4	0.207	0.075	2.760	0.006	0.060	0.354
San Jose 5	0.235	0.075	3.150	0.002	0.089	0.382
San Jose 6	0.873	0.073	11.980	<0.001	0.730	1.016
San Jose 7	-0.178	0.088	-2.030	0.043	-0.351	-0.006

Constant	1341.770	576.361	2.330	0.020	211.972	2471.569

4.1.3. Discussion

Table 10 presents all the regressions for the EnergySage Data for ease of comparison (excludes installer regressions). The EnergySage data is highly linear and shows a strong correlation with the system size variable. The strength of that correlation is powerful enough that using the Size variable alone is sufficient to serve as a predictive model, even compared to model specifications that include more detail and have other statistically significant coefficients. For the purposes of the forthcoming PV LCC tool (Present Value of Photovoltaics – PV^2), quote data from EnergySage can rely on system size only for prediction of default cost estimates for homeowners. This also prevents overfitting by focusing on installers that may not exist at future times.

Other observations can be made regarding the significant coefficients in the model. Using the specification model, no significant improvement to adjusted R^2 or MSPE is achieved, even when using all significant variables related to technology. The installer model does improve adjusted R^2 and MSPE relative to the Size only model. There are a few hypotheses as to why the installer model accounts for more variability. First, installers may be consistently using the same modules and inverters, thus the installer variable may be proxying for module, inverter and quality, while also incorporating other non-technology costs specific to the installer (e.g., operational overhead). A clearing house also may be more competitive than other solar PV markets, causing installers to price match in order to attract buyers. Lower soft costs may also be a factor. It should be noted that the improvements in adjusted R^2 and MSPE are not proof of superiority of the installer model relative to the specification model but do suggest that such a relationship may be worth further investigation through more complex analysis with a more comprehensive dataset. Last, there are multiple significant predictors that using the Size only model ignores. Inverter type, quality, and city are statistically significant in solar PV quote models. Although these do not translate to improvements in predictive power due to the large portion of the quoted cost explained by system size as well as the inherent noise in the data, the trends they represent are real and worth considering. The significant difference in the City coefficients is especially interesting because it shows statistically what is generally accepted, that locality has a significant impact on PV quotes, and suggests further research is needed in more refined markets.

Table 10. EnergySage Regression Coefficients (Dark Yellow, p <0.010, Medium Yellow, p < 0.050, Light Yellow, p < 0.100)

Equation	1	2	3a	3b	4	6
Adjusted R^2	0.948	0.948	0.948	0.948	0.949	0.958
Size	2.788	2.790	2.840	2.840	2.784	3.049
Tier						
Standard		-251.037			-377.268	
Tier#Size						
Standard		-0.0615			-0.051	
Inv						
String			1060.241			
Inv#Size						
String			-0.0640			
InvOpt						
String No Opt				1740.260	3409.467	
String Opt				1055.893	650.080	
InvOpt##Size						
String No Opt				-0.100	-0.376	
String Opt				-0.064	-0.006	
City						
Los Angeles						-312.693
San Diego						206.677
San Francisco						576.953
San Jose						-34.950
City#Size						
Los Angeles						-0.269
San Diego						-0.290
San Francisco						0.180
San Jose						0.025
Constant	1523.913	1612.984	668.922	668.922	1162.598	957.548

4.2. Tracking the Sun Regressions

The TTS data analysis follows a similar pattern to the EnergySage analysis with a few differences. First the dependent variable being regressed is the installed price of the system, as opposed to the quote. Therefore, TTS data points are all real system installations, as opposed to quotes for potential systems. As quoted prices and installed prices for the same system can differ due to changes to system designs, delays, permitting issues, or other unexpected problems, installed prices are far less certain and more variable. The TTS data also is more detailed in terms of technical specifications than the EnergySage data, though it lacks a quality variable. Figure 3 shows another difference in the two data sets, as the TTS data is far more dispersed than the EnergySage quotes shown in Figure 2. Note that all regressions are done on the same set of 2514 observations.

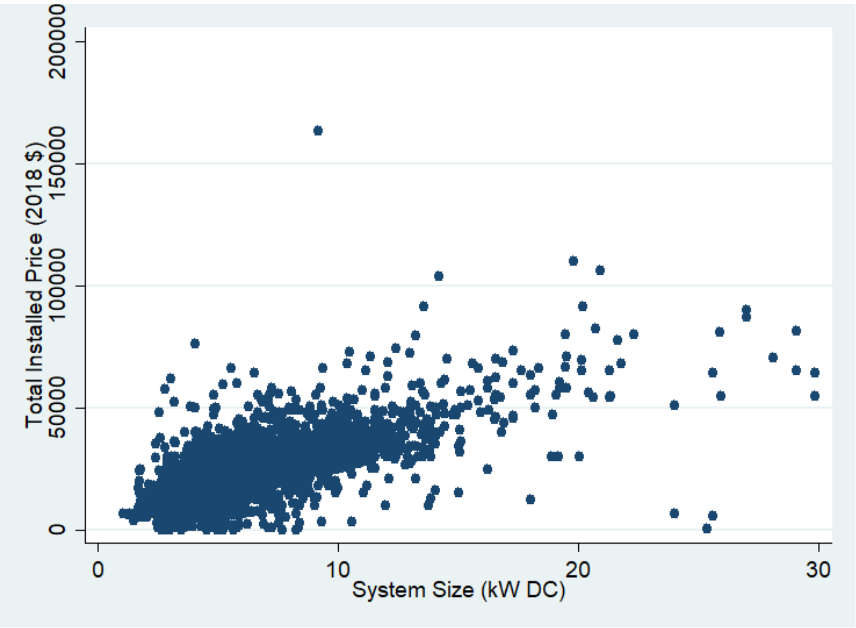

Figure 3. Plot of Total Installed Price against System Size for the Filtered TTS Data

Heteroskedasticity isn't as evident as the EnergySage data due to the dispersed nature of the TTS data, however a Breusch-Pagan test confirms heteroskedasticity between Size and Price. Considering that Size is required for all regressions that follow, all regressions utilize robust standard errors.

Regressing Price on Size according to Equation 9 yields The adjusted R2 for the model is 0.6950 and the MSPE is 6.4126E7. Unlike the EnergySage data that shows a high degree of correlation between the Quote and Size variables, the TTS data shows moderate correlation and a much higher prediction error. This outcome is likely due to the idiosyncrasies involved in working on a specific project as opposed to dealing with a quote which is often more standardized and potentially optimistic. The marginal cost is \$0.60/WDC (22 %) higher at \$3.39 per WDC compared to \$2.79 per WDC relative to the EnergySage estimate, implying higher reported installed costs than the quoted installed costs from the online platform. However, a true comparison isn't feasible due to the natural lag between obtaining a quote and deciding on and finishing installation of a system. Additionally, the EnergySage data may account for a specific subset of the overall market due to its online platform nature. As before, a specification-based model and an installer-based model are developed from the available variables in the TTS data set.

Table 11 (marginal cost in $\$/\mathrm{kW}_{\mathrm{DC}}$).

$$Price = \beta_1 * Size + \beta_0 + \epsilon_0 \tag{9}$$

The adjusted R^2 for the model is 0.6950 and the MSPE is 6.4126E7. Unlike the EnergySage data that shows a high degree of correlation between the Quote and Size variables, the TTS data shows moderate correlation and a much higher prediction error. This outcome is likely due to the idiosyncrasies involved in working on a specific project as opposed to dealing with a quote which is often more standardized and potentially optimistic. The marginal cost is \$0.60/W$_{DC}$ (22 %) higher at \$3.39 per W$_{DC}$ compared to \$2.79 per W$_{DC}$ relative to the EnergySage estimate, implying higher reported installed costs than the quoted installed costs from the online platform. However, a true comparison isn't feasible due to the natural lag between obtaining a quote and deciding on and finishing installation of a system. Additionally, the EnergySage data may account for a specific subset of the overall market due to its online platform nature. As before, a specification-based model and an installer-based model are developed from the available variables in the TTS data set.

Table 11. Equation 9 Regression Results

	Robust Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
Size	3.388	0.070	48.12	<0.001	3.250	3.526
Constant	2.071	0.367	5.65	<0.001	1.352	2.790

4.2.1. Specification Models

Module efficiency is used to proxy for the quality variable defined in the EnergySage data (see Equation 10 for the model specification).

$$Price = \beta_1 * Size + \beta_2 * Eff\#Size + \beta_3 * Eff + \beta_0 + \epsilon_0 \qquad (10)$$

Table 12 provides the results for the Efficiency regression. Of note is the fact that the Efficiency and Size interaction is not significant while the Efficiency fixed cost variable is significant.

Table 12. Equation 10 Regression Results

	Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
Size	2.956	0.936	3.16	0.002	1.122	4.791
Eff	57940.040	22747.120	2.55	0.011	13335.000	102545.100
Eff#Size	1.898	4.660	0.41	0.684	-7.240	11.035
Constant	-8994.140	4494.123	-2.00	0.045	-17806.710	-181.571

Decomposing the regression into just the Efficiency variable produces Equation 11.

$$Price = \beta_1 * Size + \beta_2 * Eff + \beta_0 + \epsilon_0 \qquad (11)$$

Table 13 gives the results. All coefficients are significant, indicating that Eff is a significant predictor. Efficiency is measured in percentage in the TTS data set, so interpreting the coefficient is not as straightforward since an increase of efficiency of one unit would result in a 100 % efficiency, which is not possible. One could naively say however that a 100 % efficient system would add roughly \$62 000 to the price of the system. Because efficiency is independent of system size in the above model this increase would be a flat rate.

Interestingly, the fixed cost is now negative and statistically significant. This is likely a result of the efficiency representing a fixed value increase independent of the Size variable, causing the constant to adjust to account for it. Furthermore, there are no systems below roughly 16 % efficiency, so the data has no points near the origin in relation to the efficiency axis. In this case the constant cannot be readily interpreted as anything more than an adjustment to minimize the loss function.

Table 13. Equation 11 Regression Results

	Robust Coef.	Robust Std. Err.	T	P>t	[95% Conf. Interval]	
Size	3.324	0.073	45.260	<0.001	3.180	3.468
Eff	61867.760	9201.793	6.720	<0.001	43824.000	79911.510
Constant	-9596.575	1697.279	-5.650	<0.001	-12924.760	-6268.388

In terms of predictive power, the addition of efficiency is insignificant, with the adjusted R^2 increasing to only 0.7027 and the MSPE becoming 6.3224E7 relative to the Equation 9 regression. For predictive purposes the efficiency variable can be omitted.

The Efficiency regression with only the interaction variable per Equation 12 is used to evaluate the marginal effects of efficiency that may be hidden by the fixed cost effects from the previous regression.

$$Price = \beta_1 * Size + \beta_2 * Eff\#Size + \beta_0 + \epsilon_0 \tag{12}$$

The results of Equation 12 are presented in Table 14. If only the interaction is included, then it becomes significant. Using the Akaike and Bayesian Information Criteria, the Eff model is not distinguishable from the interaction model (Efficiency AIC is 51742.71, BIC is 51760.24, for the interaction model AIC is 51746.97, BIC is 51764.5). Given their near identical nature, expert judgement can be used to guide model development. Since efficiency is directly associated with the solar PV panels, it is expected to directly impact the marginal cost in practice and the predictive power is statistically indistinguishable, Equation 12 is selected.

Table 14. Equation 12 Regression Results

	Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
Size	1.425	0.437	3.26	0.001	0.568	2.283
Eff#Size	9.864.615	2.010	4.91	<0.001	5922.523	13806.710
Constant	2111.593	377.1237	5.60	<0.001	1372.088	2851.099

Looking next at inverter type the regression becomes Equation 13:

$$Price = \beta_1 * Size + \beta_2 * Size\#i.Inv + \beta_3 * i.Inv + \beta_0 + \epsilon_0 \tag{13}$$

Table 15 summarizes the regression results. The regression suggests that string inverters are less expensive than microinverters by $1850 of fixed cost with no statistical difference in marginal costs. This seems counterintuitive given that system size should determine the inverter size and thus the total price. One explanation could simply be that both micro and string inverters are sized to roughly the same capacity, therefore the marginal effect washes out leaving only the difference in installation cost.

Table 15. Equation 13 Regression Results

	Robust Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
Size	3.433	0.085	40.410	<0.001	3.266	3.600
Inv						
String	-1849.644	752.892	-2.460	0.014	-3325.986	-373.301
Inv#Size						
String	-0.039	0.142	-0.270	0.786	-0.319	0.242
Constant	2921.388	432.941	6.750	<0.001	2072.435	3770.340

Regardless, the Size and Inverter Type interaction can be removed from Equation 13 resulting in the following form for the inverter type model in Equation 14.

$$Price = \beta_1 * Size + \beta_2 * i.Inv + \beta_0 + \epsilon_0 \tag{14}$$

In terms of additional predictive power, the Inverter Type variable is negligible with an adjusted R^2 of 0.7031 and an MSPE of 6.4899E7.

Adding the DC optimizer to the regression works the same way as for the EnergySage data. Regression values are found in Table 16. The results are similar to the EnergySage results with the exception that there is no significant difference between the microinverter marginal cost and either of the string inverter marginal costs. The optimizer in this case has a lower fixed cost than the microinverter as opposed to the greater price in the EnergySage data. Constraint should be used when comparing the results as the lag between quoted systems and installed systems may give sufficient time for price trends to change. The adjusted R^2 and MSPE are 0.7057 and 6.4928E7, respectively, and represent a statistically insignificant change from the Size only model.

Table 16. Equation 14 Regression Results with DC Optimizer

	Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
Size	3.433	0.085	40.390	<0.001	3.266	3.600
InvOpt						
String No Opt	-2268.463	2960.712	-0.770	0.444	-8074.112	3537.186
String Opt	-1529.150	763.515	-2.000	0.045	-3026.323	-31.976
InvOpt##Size						
String No Opt	-0.591	0.613	-0.960	0.335	-1.793	0.611
String Opt	-0.059	0.145	-0.400	0.686	-0.342	0.225
Constant	2921.388	433.111	6.750	<0.001	2072.101	3770.674

The last specification regression seeks to find a combination of the above specification models that maintains significant coefficients. The ultimate form is given in Equation 15.

$$Price = \beta_1 * Size + \beta_2 * i.Inv + \beta_3 * Eff + \beta_0 + \epsilon_0 \tag{15}$$

The regression results are found Table 17. All predictors are significant, and the same basic relationships hold in the aggregate model that existed in the piece-wise models. Higher efficiency panels are more expensive ($0.06/W per 1% in rated efficiency) and systems with microinverters are also more expensive ($4400 more than with string inverters and $1413 more than with optimizers). Predictive power is, again, not significantly increased (adjusted R^2 and MSPE are 0.7087 and 6.3426E7, respectively).

Table 17. Equation 15 Regression Results

	Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
Size	2.130	0.470	4.53	<0.001	1.207	3.053
InvOpt						
String No Opt	-4400.378	873.9334	-5.04	<0.001	-6114.071	-2686.686
String Opt	-1412	268.8247	-5.26	<0.001	-1939.898	-885.6232
Eff#Size	6.323	2.173	2.91	0.004	2.063	10.584
Constant	3033.642	366.6197	8.27	<0.001	2314.739	3752.545

Similar to the EnergySage specification model, the benefit of this model is that is allows for comparison of different system configurations. Table 18 below shows the estimated fixed cost and marginal cost based on the different configuration options. Let's compare the following: high efficiency (20 %) panels with microinverters, standard efficiency (18%) panels with string inverter and optimizers, and standard efficiency panel with sting inverter. Additional fixed costs for different inverter types may represent different installation techniques, however the cause of the difference is beyond the scope of the current paper.

Table 18. Fixed Cost and Marginal Cost by System Specification

Inverter	**Fixed Cost**
String	-1367
Opt	1621
Micro	3034

Efficiency	**Marg Cost**
16 %	3.14
18 %	3.27
20 %	3.40

Assuming a 10.0 kW system, the installed costs are estimated at $31 320, $34 307, and $36 985, respectively. The $5665 difference (Roughly 15 % to 20 % of total installed costs) across these systems would not be captured in the Size only model that would project the costs to be the same for all systems.

4.2.2. Installer Models

The same three basic regressions in Equations 5 through 7 are run for the TTS data, except using Price in lieu of Quote. All regressions with city-installer interactions are limited to the top 10 installers per city except for Los Angeles and San Francisco, where an insufficient number of city-installer groups with enough data to generate significant results were available (in total 46 % of all installations after filtering out systems per Section 3.2 and removing any systems with missing data). There is a small amount of overlap in installers between cities, reducing the total number of installers in the model further. As such 33 installers are represented instead of 50. These cities are limited to four and three installer groups, respectively.

$$Price = \beta_1 * Size + \beta_2 * Size\#i.Inst + \beta_3 * i.Inst + \beta_0 + \epsilon_0 \tag{16}$$

$$Price = \beta_1 * Size + \beta_2 * Size\#i.City + \beta_3 * i.City + \beta_0 + \epsilon_0 \tag{17}$$

$$Price = \beta_1 * Size + \beta_2 * Size\#i.CityInst + \beta_3 * i.CityInst + \beta_0 + \epsilon_0 \tag{18}$$

Looking at the installer model first (Table 19). The marginal cost of solar PV sees the most significant variables (21 of 33) relative to the base installer with variations of marginal costs from -$2.81/W to $2.00/W. Only five installers realize significant differences in fixed costs with a range from -$10 868 to $9558. Several installers (12, 15, and 18) have statistically significant differences in both fixed costs and marginal costs. In these three cases, the installers realize much higher fixed costs and lower marginal costs, which could be due to different cost structures (capacity for wholesale purchases) in those installers or artifacts of the available data from those installers.

Table 19. Equation 16 Regression Results

	Coef.	Robust Std. Err.	t	P>t	[95% Conf.	Interval]
Size	4.407	0.195	22.590	<0.001	4.024	4.789
Inst						
2	62.702	2259.537	0.030	0.978	-4368.099	4493.503
3	246.859	2184.325	0.110	0.910	-4036.458	4530.176
4	4807.211	3839.487	1.250	0.211	-2721.768	12336.190
5	-6441.186	5542.720	-1.160	0.245	-17310.090	4427.720
6	273.806	2318.013	0.120	0.906	-4271.665	4819.276
7	5456.836	3933.097	1.390	0.165	-2255.706	13169.380
8	2147.228	4619.898	0.460	0.642	-6912.085	11206.540
9	-5481.471	4129.653	-1.330	0.185	-13579.450	2616.505
10	6799.491	6322.469	1.080	0.282	-5598.452	19197.430
11	10291.440	5411.926	1.900	0.057	-320.989	20903.860
12	8204.946	3320.190	2.470	0.014	1694.275	14715.620
13	4214.172	3162.438	1.330	0.183	-1987.159	10415.500
14	-808.555	2252.335	-0.360	0.720	-5225.235	3608.125
15	9558.215	4540.583	2.110	0.035	654.434	18462.000
16	7443.541	2526.930	2.950	0.003	2488.400	12398.680
17	-1041.680	2801.964	-0.370	0.710	-6536.145	4452.785

18	685.326	4459.628	0.150	0.878	-8059.709	9430.360
19	-2299.057	3144.813	-0.730	0.465	-8465.826	3867.712
20	3475.562	3229.732	1.080	0.282	-2857.728	9808.852
21	1715.257	2595.028	0.660	0.509	-3373.421	6803.935
22	-841.994	2244.986	-0.380	0.708	-5244.262	3560.274
23	3554.371	3934.717	0.900	0.366	-4161.347	11270.090
24	3789.344	2805.629	1.350	0.177	-1712.308	9290.995
25	1003.017	2353.253	0.430	0.670	-3611.555	5617.589
26	1675.481	2267.415	0.740	0.460	-2770.769	6121.730
27	3341.695	2586.524	1.290	0.196	-1730.306	8413.696
28	2355.239	2382.125	0.990	0.323	-2315.949	7026.428
29	-335.929	2305.281	-0.150	0.884	-4856.433	4184.575
30	-10868.080	3365.610	-3.230	0.001	-17467.810	-4268.340
31	1772.419	3403.191	0.520	0.603	-4901.012	8445.850
32	-467.514	2296.000	-0.200	0.839	-4969.816	4034.789
33	8645.316	3089.603	2.800	0.005	2586.810	14703.820
Inst#Size						
2	-1.071	0.221	-4.830	<0.001	-1.506	-0.636
3	-1.669	0.229	-7.290	<0.001	-2.119	-1.220
4	-1.376	0.430	-3.200	0.001	-2.219	-0.532
5	0.849	1.017	0.840	0.404	-1.145	2.844
6	-0.515	0.301	-1.710	0.088	-1.107	0.076
7	-1.687	0.361	-4.660	<0.001	-2.397	-0.978
8	1.294	0.431	3.000	0.003	0.448	2.139
9	-0.871	0.840	-1.040	0.300	-2.519	0.776
10	-0.087	1.181	-0.070	0.941	-2.403	2.229
11	-1.097	1.269	-0.860	0.387	-3.587	1.392
12	-2.805	0.632	-4.440	<0.001	-4.044	-1.565
13	-1.638	0.379	-4.320	<0.001	-2.382	-0.894
14	-0.023	0.259	-0.090	0.927	-0.533	0.485
15	-2.100	0.399	-5.260	<0.001	-2.883	-1.317
16	-0.382	0.316	-1.210	0.226	-1.002	0.237
17	-0.412	0.362	-1.140	0.255	-1.123	0.297
18	2.000	0.851	2.350	0.019	0.330	3.670
19	-0.408	0.278	-1.470	0.142	-0.953	0.136
20	-1.892	0.414	-4.560	<0.001	-2.705	-1.079
21	-2.147	0.360	-5.950	<0.001	-2.855	-1.439
22	-0.669	0.226	-2.950	0.003	-1.113	-0.224
23	-1.953	0.416	-4.690	<0.001	-2.770	-1.136

24	-1.187	0.267	-4.430	<0.001	-1.712	-0.662
25	-0.987	0.237	-4.160	<0.001	-1.452	-0.521
26	-1.198	0.217	-5.510	<0.001	-1.624	-0.771
27	-0.930	0.319	-2.910	0.004	-1.557	-0.303
28	-1.502	0.260	-5.760	<0.001	-2.013	-0.991
29	0.032	0.269	0.120	0.905	-0.496	0.560
30	0.442	0.385	1.150	0.251	-0.313	1.198
31	-1.322	0.300	-4.410	<0.001	-1.911	-0.734
32	-1.383	0.208	-6.630	<0.001	-1.793	-0.974
33	-2.300	0.390	-5.890	<0.001	-3.066	-1.534
Constant	378.347	2169.401	0.170	0.862	-3875.705	4632.398

The negative constants are possibly due to wide scatter for some installers or too few points near the origin to be able to meaningfully interpret behavior of the model. The EV for the model is significantly increased; however, the PE is significantly increased for this model (adjusted R^2 and MSPE are 0.8233 and 8.6218E7, respectively). The increase in PE is likely due to the smaller sample size for each Inst group. Also, the prediction intervals overlap between the Size only model and this model for nearly every data point. The installer model does show higher total installed prices for roughly 75 % of its predictions. While there is no statistical justification using the current data to say this bias is significant, should more data become available this potential bias should be investigated further.

Equation 17's regression results are found in Table 20. Statistically significant differences exist at the city level. Using Fresno as the base city, San Francisco and San Jose have statistically significant increases in fixed cost (around $5000), while Los Angeles and San Diego are statistically the same. All cities have a significant increase in marginal cost (range of $0.48/W to $1.31/W) except for San Jose, which is statistically identical. Adjusted R^2 and MSPE are not significantly improved (0.7372 and 6.5263E7, respectively).

Table 20. Equation 17 Regression Results

	Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
Size	3.139	0.065	47.960	<0.001	3.011	3.267
City						
Los Angeles	-2678.630	1843.921	-1.450	0.146	-6294.397	937.137
San Diego	-1063.674	689.275	-1.540	0.123	-2415.281	287.933
San Francisco	4880.999	2095.650	2.330	0.020	771.613	8990.384
San Jose	5456.736	1427.892	3.820	<0.001	2656.765	8256.706
City#Size						
Los Angeles	1.099	0.341	3.220	0.001	0.431	1.768
San Diego	0.484	0.129	3.730	<0.001	0.229	0.738
San Francisco	1.311	0.445	2.950	0.003	0.438	2.184
San Jose	0.202	0.233	0.870	0.387	-0.255	0.660
Constant	1427.348	407.115	3.510	<0.001	629.031	2225.664

Last, the city-installer group model is presented in Table 21. Significant differences appear when using the model in Equation 16, but the improvement over the Size only model is negligible in terms of MSPE (6.4348E7) and over 99 % of prediction intervals overlapping between the two but shows a significant increase in adjusted R^2 (0.8242). The statistically significant variation of $21 160 (-$9824 to $11 335) in the city-installer fixed costs and $4.85/W (-$2.81/W to $2.00/W) in marginal costs are like those found in the installer only model ($20 426 and $4.81/W). As before the constant becomes negative for a small number of groups. Based on the results, using the Size only model is justified for predictive purposes.

Table 21. Equation 18 Regression Results

	Coef.	Robust Std. Err.	t	P>t	[95% Conf.	Interval]
Size	3.445	0.117	29.390	<0.001	3.215	3.675
CityInst						
Fresno 2	1290.878	854.042	1.510	0.131	-383.843	2965.600
Fresno 3	6500.855	3385.622	1.920	0.055	-138.136	13139.850
Fresno 4	10602.230	4077.616	2.600	0.009	2606.287	18598.180
Fresno 5	-1255.038	2421.723	-0.520	0.604	-6003.882	3493.807
Fresno 6	4598.391	3387.510	1.360	0.175	-2044.302	11241.080
Fresno 7	273.349	1162.185	0.240	0.814	-2005.622	2552.320
Fresno 8	2816.439	2749.936	1.020	0.306	-2576.012	8208.889
Fresno 9	576.506	1109.668	0.520	0.603	-1599.484	2752.496
Fresno 10	9689.335	2349.343	4.120	<0.001	5082.423	14296.250
Los Angeles 1	1044.020	2320.761	0.450	0.653	-3506.845	5594.885

Los Angeles 2	1317.825	1154.673	1.140	0.254	-946.415	3582.066
Los Angeles 3	9248.966	2646.180	3.500	<0.001	4059.974	14437.960
Los Angeles 4	-9824.058	2703.134	-3.630	<0.001	-15124.730	-4523.382
San Diego 1	-4437.451	3612.813	-1.230	0.219	-11521.950	2647.047
San Diego 2	235.465	1015.930	0.230	0.817	-1756.710	2227.640
San Diego 3	2.340	1954.280	0.000	0.999	-3829.879	3834.558
San Diego 4	2759.277	1642.777	1.680	0.093	-462.105	5980.659
San Diego 5	202.026	999.476	0.200	0.840	-1757.883	2161.934
San Diego 6	2047.037	1224.100	1.670	0.095	-353.346	4447.419
San Diego 7	2719.500	1049.043	2.590	0.010	662.393	4776.608
San Diego 8	4385.714	1629.265	2.690	0.007	1190.829	7580.600
San Diego 9	3399.259	1278.902	2.660	0.008	891.414	5907.104
San Diego 10	417.393	1247.206	0.330	0.738	-2028.299	2863.086
San Francisco 1	7843.510	6003.928	1.310	0.192	-3929.812	19616.830
San Francisco 2	8487.561	1532.571	5.540	<0.001	5482.285	11492.840
San Francisco 3	2894.372	5503.406	0.530	0.599	-7897.458	13686.200
San Jose 1	1649.182	1661.615	0.990	0.321	-1609.141	4907.505
San Jose 2	5851.230	3276.043	1.790	0.074	-572.883	12275.340
San Jose 3	-5397.167	5173.485	-1.040	0.297	-15542.040	4747.709
San Jose 4	3191.248	4166.042	0.770	0.444	-4978.096	11360.590
San Jose 5	11335.460	5032.637	2.250	0.024	1466.775	21204.140
San Jose 6	5258.192	2444.641	2.150	0.032	464.405	10051.980
San Jose 7	4021.253	4640.380	0.870	0.386	-5078.238	13120.740
San Jose 8	4519.582	2531.371	1.790	0.074	-444.276	9483.439
San Jose 9	4833.363	1959.548	2.470	0.014	990.814	8675.912
San Jose 10	2266.026	1003.012	2.260	0.024	299.183	4232.868
CityInst#Size						
Fresno 2	-0.707	0.168	-4.210	<0.001	-1.037	-0.378
Fresno 3	-0.725	0.326	-2.220	0.026	-1.367	-0.084
Fresno 4	-1.138	0.368	-3.090	0.002	-1.860	-0.417
Fresno 5	0.553	0.230	2.400	0.016	0.101	1.005
Fresno 6	-0.991	0.386	-2.560	0.010	-1.750	-0.233
Fresno 7	1.057	0.265	3.990	<0.001	0.537	1.578
Fresno 8	-0.360	0.256	-1.400	0.160	-0.864	0.142
Fresno 9	-0.421	0.138	-3.040	0.002	-0.694	-0.149
Fresno 10	-1.338	0.358	-3.730	<0.001	-2.041	-0.635
Los Angeles 1	0.961	0.227	4.220	<0.001	0.515	1.408
Los Angeles 2	0.446	0.258	1.720	0.085	-0.061	0.953
Los Angeles 3	-1.843	0.613	-3.000	0.003	-3.046	-0.640

Los Angeles 4	1.404	0.352	3.980	<0.001	0.712	2.096
San Diego 1	0.090	0.827	0.110	0.913	-1.531	1.712
San Diego 2	0.938	0.208	4.510	<0.001	0.530	1.346
San Diego 3	0.549	0.327	1.680	0.094	-0.093	1.191
San Diego 4	-1.185	0.325	-3.640	<0.001	-1.824	-0.546
San Diego 5	0.292	0.164	1.780	0.075	-0.030	0.615
San Diego 6	-0.025	0.179	-0.140	0.888	-0.376	0.325
San Diego 7	-0.236	0.151	-1.560	0.119	-0.533	0.060
San Diego 8	0.031	0.279	0.110	0.911	-0.516	0.579
San Diego 9	-0.540	0.209	-2.580	0.010	-0.950	-0.130
San Diego 10	1.080	0.260	4.150	<0.001	0.569	1.591
San Francisco 1	0.874	1.173	0.750	0.456	-1.425	3.175
San Francisco 2	0.579	0.275	2.100	0.035	0.039	1.119
San Francisco 3	1.841	1.113	1.650	0.098	-0.341	4.025
San Jose 1	-0.163	0.326	-0.500	0.617	-0.803	0.476
San Jose 2	-0.414	0.401	-1.030	0.303	-1.201	0.373
San Jose 3	1.811	1.007	1.800	0.072	-0.162	3.786
San Jose 4	2.256	0.402	5.600	<0.001	1.466	3.045
San Jose 5	-0.135	1.262	-0.110	0.914	-2.611	2.339
San Jose 6	-0.676	0.346	-1.950	0.051	-1.355	0.003
San Jose 7	3.010	0.985	3.060	0.002	1.078	4.942
San Jose 8	-0.930	0.384	-2.420	0.016	-1.684	-0.175
San Jose 9	-0.225	0.218	-1.030	0.301	-0.653	0.202
San Jose 10	0.685	0.142	4.830	<0.001	0.407	0.964
Constant	-665.673	814.984	-0.820	0.414	-2263.804	932.459

4.2.3. Discussion

Table 20 presents the reduced regressions for the TTS data (excluding installer regressions) for ease of comparison. Looking at the models from a prediction perspective, any pricing tool may use the Size only model without losing any statistically significant gains from other significant predictors if using the TTS data. Thus, the form in Equation 9 is satisfactory. However, the multiple models show that there are significant predictors that need to be accounted for if looking from an explanatory perspective. The module efficiency acts as an adjustment to marginal cost, while inverter type and city (including the DC optimizer) play a role in determining the ultimate price of solar both in terms of fixed and marginal costs. There are differences across cities, showing that market-specific estimates are appropriate. Installer, as with the EnergySage data, explains the most variance on its own, likely due to the aforementioned proxying of technology coupled with the implicit inclusion of installer-specific fixed costs. However, it fails to increase actual predictive power (relative to MSPE or prediction intervals) in a meaningful way.

Table 22. Summary of Reduced TTS Expressions

Equation	9	12	13	14	15	17
Adjusted R^2	0.6950	0.7027	0.7031	0.7057	0.7087	0.7327
Size	3387.833	1425.346	3433.154	3433.154	2130.432	3139.614
Eff#Size		9864.615			6323.456	
Inv						
String			-1849.644			
Inv#Size						
String			-38.879			
InvOpt						
String No Opt				-2268.463	-4400.378	
String Opt				-1529.15	-1412.761	
InvOpt#Size						
String No Opt				-590.772		
String Opt				-58.506		
City						
Los Angeles						-2678.63
San Diego						-1063.674
San Francisco						4880.999
San Jose						5456.736
City#Size						
Los Angeles						1099.926
San Diego						484.188
San Francisco						1311.313
San Jose						202.225
Constant	2071.054	2111.593	2921.388	2921.388	3033.642	1427.348

5. Conclusion

Total solar PV installations in the U.S. continue to increase significantly each year. Policy decisions and the nature of solar markets continue to shift; however, it is likely that the price of solar will continue to decrease in the near term. Given the increasing market and more competition in installations, it is beneficial to have a greater understanding in the driving factors in solar PV pricing, as well as models to help perspective buyers and sellers to obtain estimates for the cost of installations.

At present the most common model for solar PV pricing is solely based on marginal costs by the size of the solar PV system. The work in Webb et al. [5] shows that this is likely impacting estimates of solar PV pricing by ignoring the fixed cost component. In an examination of two data sets for California for installations and quotes for 2018, some key findings emerge. First, for the data used, system size with a fixed cost component is a sufficient predictor. While it does not explain the most variation in the data, the model produces estimates that are statistically indistinguishable from more complicated models. Whether this holds for all data sets is unknown, however the process for making such a determination is laid out here.

Second, using system size by itself glosses over other significant predictors by attempting to "bake" them into the model. The inverter technology, quality (or efficiency) of the panels, and the city all are important in determining the ultimate price of a quote or installed price for a system and may not show up as marginal impacts. Also, an installer regression model, with system size, manages to capture more of the variation than using the specifications by themselves. This indicates that installer is a possible proxy for the specification variables, as well as incorporating pricing impacts not included in the specifications available in the data. While all of this is not entirely unsurprising, having the statistical basis informs decisions on the development of predictive and explanatory models going forward, as well as other areas of vital data collection and research.

The current work is meant to serve as an initial probe into the data sets using rudimentary methods. Future work could include multiple topics. A deeper dive into the spatial component of pricing, utilizing ZIP code groupings and the physical location of installers, may provide better insight into market competition, it's pricing impacts, and how markets develop if sufficient historical data is provided. Linking the EnergySage and TTS data sets would provide the opportunity to see the rate at which quotes become installed systems, and how quotes compare with installed prices. Doing so would require additional data not available for the current paper. The use of finer time periods, time-series, seasonal and autoregressive models could also check for lagged effects or if solar PV pricing varies at time frames less than a year, provided sufficient data exists to reduce data from the yearly aggregate.

Other possibilities include looking at more complex models to examine if their prediction power is better. Given the large number of variables in the data sets, OLS quickly becomes limited, however the use of lasso regression or other machine learning techniques could incorporate more variables. An artificial neural network could be developed for instance, that would be able to take the specific panel designation and update predictions using it. A classification model could also be created to determine if it is possible to predict an installer using only system specifications. This would serve as a check of installers purchase patterns, namely if installer does serve as a proxy for specifications. With historical data available this can be traced to look at the movement of purchase decisions over time. Some of the aforementioned topics would require more data and in some cases the collection of more data than is currently done in data sets like EnergySage or the public TTS data.

References

[1] Barbose G, Darghouth N, Elmallah S, Forrester S, Kristina SH K, Millstein D, Rand J, Cotton W, Sherwood S, O'Shaughnessy E (2019) Tracking the Sun: Pricing and Design Trends for Distributed Photovoltaic Systems in the United States-2019 Edition.

[2] EnergySage (2020) EnergySage's Solar Marketplace Intel Report CY2019.

[3] Fu R, Margolis RM, Feldman DJ (2018) US Solar Photovoltaic System Cost Benchmark: Q1 2018. (National Renewable Energy Lab.(NREL), Golden, CO (United States)).

[4] O'Shaughnessy EJ , Margolis RM (2017a) The Value of Transparency in Distributed Solar PV Markets. (National Renewable Energy Lab.(NREL), Golden, CO (United States)).

[5] Webb D, Kneifel J, O'Fallon C (2020) Developing Cost Functions for Estimating Solar Photovoltaic System Installed and Life Cycle Costs Using Historical Quote Data.

[6] O'Shaughnessy EJ (2018) The Evolving Market Structure of the US Residential Solar PV Installation Industry, 2000-2016. (National Renewable Energy Lab.(NREL), Golden, CO (United States)).

[7] O'Shaughnessy E (2019) Non-monotonic effects of market concentration on prices for residential solar photovoltaics in the United States. *Energy Economics* 78:182-191.

[8] Marszal AJ , Heiselberg P (2011) Life cycle cost analysis of a multi-storey residential net zero energy building in Denmark. *Energy* 36(9):5600-5609.

[9] Leckner M , Zmeureanu R (2011) Life cycle cost and energy analysis of a Net Zero Energy House with solar combisystem. *Applied Energy* 88(1):232-241.

[10] Kannan R, Leong K, Osman R, Ho H, Tso C (2006) Life cycle assessment study of solar PV systems: An example of a 2.7 kWp distributed solar PV system in Singapore. *Solar energy* 80(5):555-563.

[11] Swift KD (2013) A comparison of the cost and financial returns for solar photovoltaic systems installed by businesses in different locations across the United States. *Renewable Energy* 57:137-143.

[12] Lang T, Gloerfeld E, Girod B (2015) Don′ t just follow the sun–A global assessment of economic performance for residential building photovoltaics. *Renewable and Sustainable Energy Reviews* 42:932-951.

[13] Farias-Rocha AP, Hassan KMK, Malimata JRR, Sánchez-Cubedo GA, Rojas-Solórzano LR (2019) Solar photovoltaic policy review and economic analysis for on-grid residential installations in the Philippines. *Journal of Cleaner Production* 223:45-56.

[14] Kouhestani FM, Byrne J, Johnson D, Spencer L, Hazendonk P, Brown B (2019) Evaluating solar energy technical and economic potential on rooftops in an urban setting: the city of Lethbridge, Canada. *International Journal of Energy and Environmental Engineering* 10(1):13-32.

[15] Kalke D, Kokkonda K, Kulkarni P (2018) Financial Analysis of Grid-tied Rooftop Solar Photovoltaic System employing Net-Metering. *2018 International Conference on Smart Electric Drives and Power System (ICSEDPS)*, (IEEE), pp 87-92.

[16] Patel AM , Singal SK (2018) LCC analysis for economic feasibility of rural electrification by hybrid energy systems. *Materials Today: Proceedings* 5(1):1556-1562.

[17] Rodríguez LR, Lissén JMS, Ramos JS, Jara EÁR, Domínguez SÁ (2016) Analysis of the economic feasibility and reduction of a building's energy consumption and emissions when integrating hybrid solar thermal/PV/micro-CHP systems. *Applied Energy* 165:828-838.

[18] Lee M, Hong T, Koo C, Kim C-J (2018) A break-even analysis and impact analysis of residential solar photovoltaic systems considering state solar incentives. *Technological and Economic Development of Economy* 24(2):358-382.

[19] Campoccia A, Dusonchet L, Telaretti E, Zizzo G (2009) Comparative analysis of different supporting measures for the production of electrical energy by solar PV and Wind systems: Four representative European cases. *Solar Energy* 83(3):287-297.

[20] Campoccia A, Dusonchet L, Telaretti E, Zizzo G (2014) An analysis of feed'in tariffs for solar PV in six representative countries of the European Union. *Solar Energy* 107:530-542.

[21] Dusonchet L , Telaretti E (2015) Comparative economic analysis of support policies for solar PV in the most representative EU countries. *Renewable and Sustainable Energy Reviews* 42:986-998.

[22] Sow A, Mehrtash M, Rousse DR, Haillot D (2019) Economic analysis of residential solar photovoltaic electricity production in Canada. *Sustainable Energy Technologies and Assessments* 33:83-94.

[23] Burns JE , Kang J-S (2012) Comparative economic analysis of supporting policies for residential solar PV in the United States: Solar Renewable Energy Credit (SREC) potential. *Energy Policy* 44:217-225.

[24] Bauner C , Crago CL (2015) Adoption of residential solar power under uncertainty: Implications for renewable energy incentives. *Energy Policy* 86:27-35.

[25] EnergySage (2019) *EnergySage Solar Photovoltaic Installation Quotation Database - Subset: California 2018*. Available at https://www.energysage.com/data/.

[26] O'Shaughnessy E , Margolis R (2017b) Using residential solar pv quote data to analyze the relationship between installer pricing and firm size. (National Renewable Energy Lab.(NREL), Golden, CO (United States)).