# Whole Building Energy Efficiency and Energy Savings Estimation: Does Smart Meter Data with Pre-screening Open up Design and Evaluation Opportunities?

*Josh L. Bode, Nexant Inc.*
*Leo Carrillo and Mangesh Basarkar, Pacific Gas and Electric*

## ABSTRACT

Whole building energy efficiency focuses on deeper, more comprehensive energy savings that, in theory, can be measured at a whole building level using smart meter data. This paper summarizes the results from an assessment of whether smart meters, with and without pre-screening of sites, can be used to accurately estimate energy savings for performance-based payments. The accuracy of savings estimates and the degree to which different screening criteria improved accuracy was estimated using 1,446 buildings. In addition, we applied 767 combinations of screening criteria to assess if pre-screening could improve accuracy of energy savings estimates by identifying and excluding hard-to-predict customers. In specific, we address five main questions: By how much does energy consumption absent energy efficiency vary from year-to-year? What are characteristics that define predictable buildings? What is the amount of error in estimated baselines when performance is measured at the individual building level? How accurately do different screening criteria identify customers that are and are not predictable? To what degree does applying pre-screening help increase the accuracy of energy savings estimates produced using smart meters? Our main conclusion is that although screening can help identify hard-to-predict customers, many customers that pass screening criteria experience disruptive changes unrelated to weather patterns or energy efficiency that affect year-to-year energy consumption. As a result, the baseline error for many customers can be larger than whole building energy savings from energy efficiency improvements. It is important to exercise caution in using energy efficiency savings estimated produced for individual sites. If they are to be used, we recommend limiting their applications to instances where customers are expected to deliver large reductions in whole building consumption.

## Introduction

The increasing availability of smart meter data has several implications for evaluation and program design. Wide scale deployment of smart meters provide additional tools with which to estimate the impact of energy efficiency, including large sample sizes, more granular data (which can be better related to occupancy and weather), faster feedback, and additional analytical techniques. It also opens new opportunities for program design and delivery, including whole building energy efficiency, better targeting of high potential customers, benchmarking of end-use efficiency and the use of performance based payments.

However, not all commercial buildings are well suited for whole building measurement. Some buildings have highly variable energy consumption, making it difficult to establish reliable baselines. As we detail later, a large share of buildings experience changes in consumption due to unobserved factors – such as changes in equipment, production, or occupancy levels. This study addressed five main research questions:

1. By how much does energy consumption absent energy efficiency vary from year-to-year?

2. What are characteristics that define predictable buildings?
3. What is the amount of error in estimated energy savings when performance is measured at the individual building level?
4. How accurately do different screening criteria identify customers that are and are not predictable?
5. To what degree does applying pre-screening help increase the accuracy of energy savings estimates produced using smart meters?

The analysis focuses exclusively on analyzing energy efficiency impacts for individual sites using a time series of electricity usage patterns before and after installation. It assumes utilities lacks information such as building occupancy levels or business schedules. This evaluation technique is known in program evaluation literature as an interrupted time series, since the introduction of treatment (energy efficiency) should lead to a change in usage patterns when it is introduced (Shadish, Cook and Campbell 2002). It is a within-subjects analysis in that it does not make use of an external control group. There is a non-trivial chance that factors other than energy efficiency influenced electricity in the same time frame, and thus may be confounded with energy efficiency. Even when differences are due to observable factors, such as weather, the results of interrupted time series can be sensitive to the model selected (e.g., was the weather relationship specified correctly?). Despite some drawbacks, it is sometimes the best available evaluation method, and it is one of the few approaches that can be used to produce customer specific energy savings estimates, which can have significant value, if accurate. The ability to link energy efficiency incentives to individual customer performance and the ability to track the amount of energy savings over time are both valuable from a program design perspective. On the other hand, some customers are also interested in tracking savings from energy efficiency upgrades to understand the return from energy efficiency investments.

The study was conducted as part of an effort to determine if performance based payments could be used in the design of whole building energy efficiency programs. While the study was conducted to assess the viability of performance based payments, it has a wide range of applications for evaluation, program design and implementation. Throughout this paper, performance refers to whether realized energy savings match, exceed, or fall short of the estimates produced prior to the installation.

Traditional energy efficiency incentives are typically paid in full when the installation of energy efficiency upgrades has been verified. Under this paradigm, performance is not estimated for individual sites and does not play a role in the amount of incentives paid to participants (since the incentives are paid in full upon verification of the installation). The logic of this approach is rooted in history. Prior to widespread adoption of smart meters, monthly billing data did not include enough observations before and after the energy efficiency upgrades to produce reliable customer specific saving estimates.

A performance based incentive approach pays part or all of the energy efficiency incentives after installation has been verified and performance has been assessed. That is, performance based payments take into account the estimated energy savings realized after the installation. The ability to develop an accurate baseline of what energy consumption would have been at that site without energy efficiency upgrades is fundamental to the concept of performance based payments. The energy baseline is the starting point to calculate energy savings, which determine the amount of incentives paid to program participants.

The remainder of this paper summarizes the methods, presents data on energy consumption patterns, discusses the characteristics that define predictable buildings, and presents the results regarding accuracy of baselines and energy savings estimates. We conclude by summarizing the key findings and their implications for program evaluation and design.

## Methodology

Figure 1 shows an overview of the mains steps in the study. We relied on a representative sample of 1,446 commercial buildings that received electric service from PG&E. A subset of those buildings, 1,168 (81%), also received natural gas service from PG&E. The sampling was stratified so one third of buildings had square footage above15,000; another third had between 5,000 and 15,000 square feet; and the remaining third had 5,000 or less square feet.

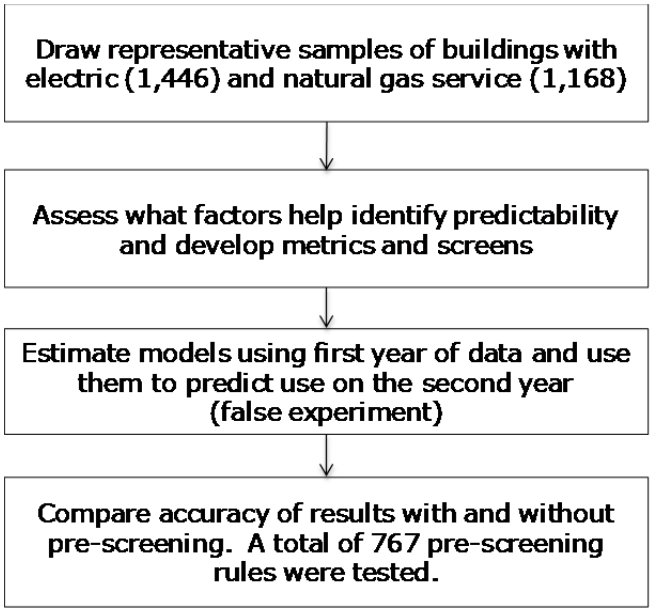



Figure 1. Method overview.

Next, we assessed the year to year variability and predictability of building energy consumption using two years of monthly billing data. The initial screens were intentionally developed using monthly billing data because the screens could be developed and applied to all of the PG&E's business accounts. Variability was defined by a standard measure, the coefficient of variation, which is simply the standard deviation divided by the average usage for a particular customer.[1] Many buildings with highly variable usage, however, also have highly predictable usage because the variation is mostly due to variation in weather conditions. The predictability metric summarized, at a high level, how predictable a customer's month to month consumption patterns were after taking into account weather. It is represented by the normalized root mean

[1] $CV = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(kWh_t - \overline{kWh})^2}}{\overline{kWh}}$

squared error, or CV RMSE.[2] A lower CV RMSE value implies that the customer's load is more predictable. The estimates of predictability were produced by running a simple regression predicting monthly usage (expressed in kWh for electricity and therms for gas) for each customer as a function of weather (cumulative cooling and heating degree hours).[3] The predictability metric is always bounded by the variability metric. That is, if none of the volatility in usage was predictable, the predictability metric would be equal to the metric for volatility; otherwise it is lower. The initial estimates of predictability and volatility were analyzed to identify and screen out customers who were not predictable. The logic was that customers with highly variable loads could be screened out from performance based payments. The third main step was implementing what is known as a placebo test or false experiment using the more granular smart meter whole building data. A false experiment is an exercise where we use the evaluation method to estimate a treatment that does not actually exist. Because there is no treatment, if a baseline is accurate, it should estimate zero energy savings due to energy efficiency since such improvements were not installed. If it does show a change, it is mistaken; other unobserved factors that led to changes in electricity usage are being mistaken for an energy efficiency impact.

In the false experiment, the baseline regression model was estimated for each individual building using approximately a year of smart meter data. We then used the regression model to predict the baseline for the second year, based on actual weather conditions at each building's location. The baseline was then used to estimate the error in the energy savings estimates. This process simulates the fact than once a customer installs energy efficiency upgrades, we are no longer able to observe what they would have consumed had they not installed them. The process also allows for a comparison of actual consumption patterns to of the baseline produced for the placebo period; thus allowing us to assess accuracy. Figure 2 illustrates the procedure.

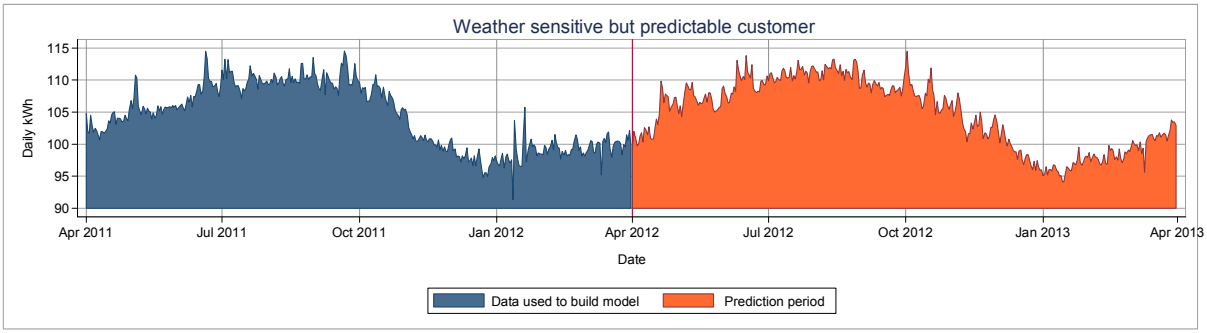



Figure 2. Illustration of out of sample placebo test.

[2] $CVRMSE = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\widehat{kWh}_t - kWh_t\right)^2}}{\overline{kWh}}$

[3] The monthly consumption regression can be summarized by the following equation:

$$consumption_{i,t} = \alpha_i + \beta_1 CDH_{i,t} + \beta_2 HDH_{i,t} + \epsilon_{i,t}$$

Where CDH and HDH reflect cooling degree hours and heating degree hours experienced over the billing period.

The regression model used to estimate the baselines is summarized by the following equation:

$$kWh_{i,t} = \alpha_i + \beta_1 DOW_t + \beta_2 holiday_t + \beta_3 CDH_{i,t} + \beta_4 HDH_{i,t} + \beta_5 daylight_{i,t} + \epsilon$$

| Term | Definition |
|---|---|
| $_{i,t}$ | Subscript i applies to each customer, while subscript t applies to each day. |
| $\alpha$ | Constant |
| $\beta_{1,2}$ | Coefficients |
| $\epsilon$ | Error term |
| $kWh$ | Daily kWh for each customer |
| $CDH$ | Cumulative daily cooling degree hours. A cooling degree hour is defined as the maximum of 0 or the temperature (in Fahrenheit) minus 70. Cumulative cooling degree hours sum these values for all hours in the day. |
| $HDH$ | Cumulative daily heating degree hours. A heating degree hour is defined as the maximum of 0 or 70 minus the temperature (in Fahrenheit). Cumulative heating degree hours sum these values for all the hours in the day. |
| $DOW$ | Dummy variable expressing whether the day is a weekend, Monday, Tuesday-Thursday or Friday. |
| $holiday$ | Dummy variable expressing whether the day is a holiday. |
| $daylight$ | Expresses the total hours of daylight on each day. |

The final step was to compare accuracy with and without pre-screening. The goal was to assess if pre-screening for hard-to-predict customers improved accuracy of energy savings estimates. In total, we assessed 767 combinations of rules based on metrics for volatility, industry, weather sensitivity, location and other factors. For brevity, results are summarized only for screening criteria that noticeably affected accuracy.

## Changes in Energy Consumption and Variability

In PG&E's territory, roughly 40% of buildings experience year to year changes in electricity consumption in excess of 10% (based on the random sample). These changes in consumption are typically unrelated to weather and thus are not captured by analyzing the relationship between weather and electricity use during the period prior to the installation of energy efficiency improvements. Figure 3 summarizes the distribution of annual changes in energy consumption.

The findings are not unique to the sample of 1,446 buildings used for this assessment. The 2012 evaluation of Southern California Edison's "10 for 10" Rebate Program, found that 36% of business customers in a control group experienced changes in electricity consumption in excess of 10% from one year to the next (George and Schellenberg 2013). An investigation into commercial building baseline modeling software embedded in energy management systems produced similar estimates of year to year changes in electricity consumption, leading its authors to conclude: "In most buildings and most years, the largest source of year-to-year change in

energy use is neither energy conservation measures nor year-to-year variation in weather, it is changes in characteristics of building operation and occupant behavior such as operating hours, thermostat settings, the number of occupants, the type of activities performed in the building, and so on (Price and Jump 2013)."

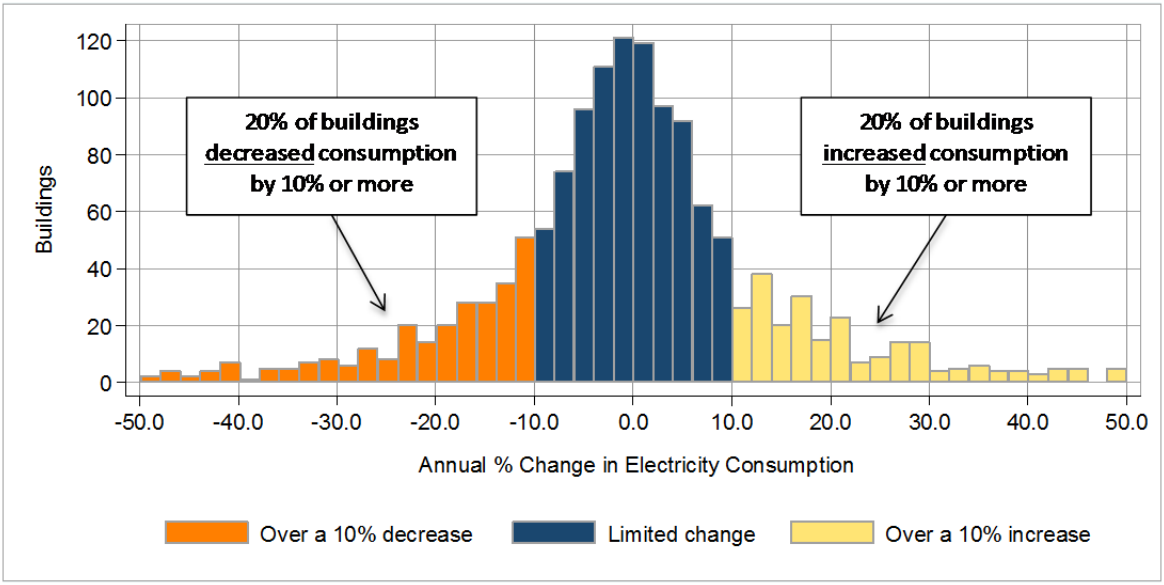



Figure 3. Distribution of year to year change in energy consumption absent energy efficiency.

Highly variable buildings can be highly predictable, especially when most of the variation in electricity use is explained by weather conditions. On the other hand, sites that appear predictable can experience disruptive changes that cannot be forecast based on prior consumption patterns. Figure 4 provides one such example. Any model that relies on the "before" usage patterns could not predict the "after" usage patterns because the abrupt change is not due to weather or energy efficiency. However, because the external change occurred in the same time frame as the energy efficiency placebo, it is quite easy for time series models to incorrectly attribute the increase in energy consumption to energy efficiency.

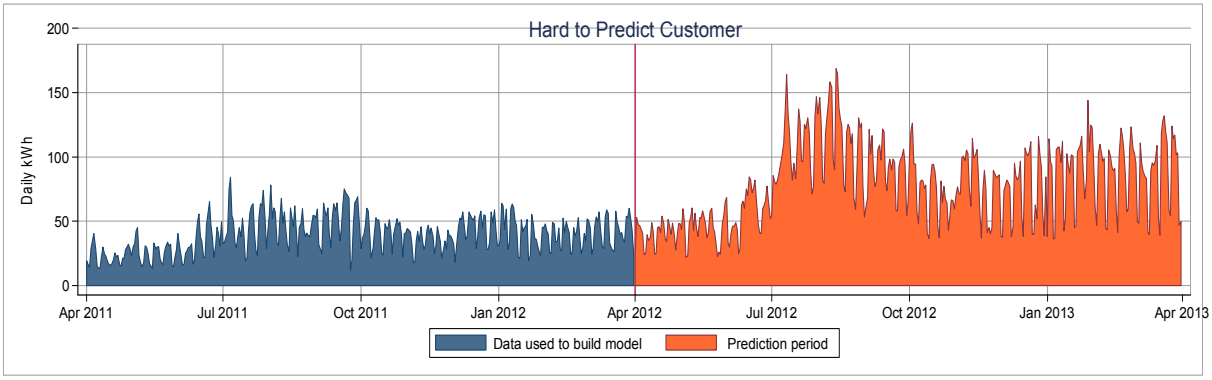
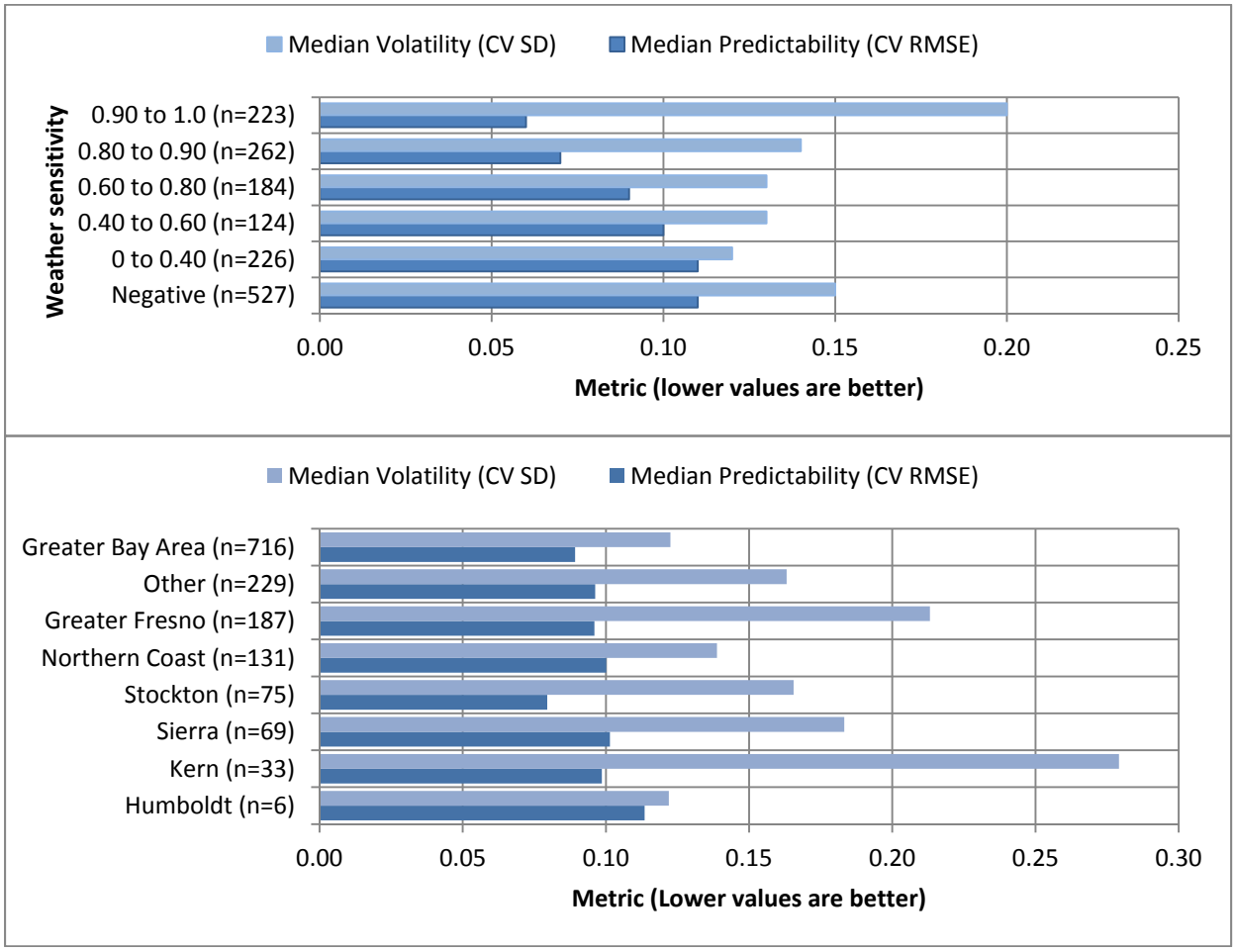



Figure 4. Example of change in consumption unrelated to weather or energy efficiency.

## Characteristics that Define Predictable Buildings

A key question is whether customer characteristics such as weather sensitivity, business type, or location are useful ways to identify and screen out hard-to-predict customers. Figure 5 summarizes the variability and predictability metrics for these customer segments. As noted earlier, the predictability metric is always lower than the variability metric since it reflects the amount of variability that is unexplained by weather.



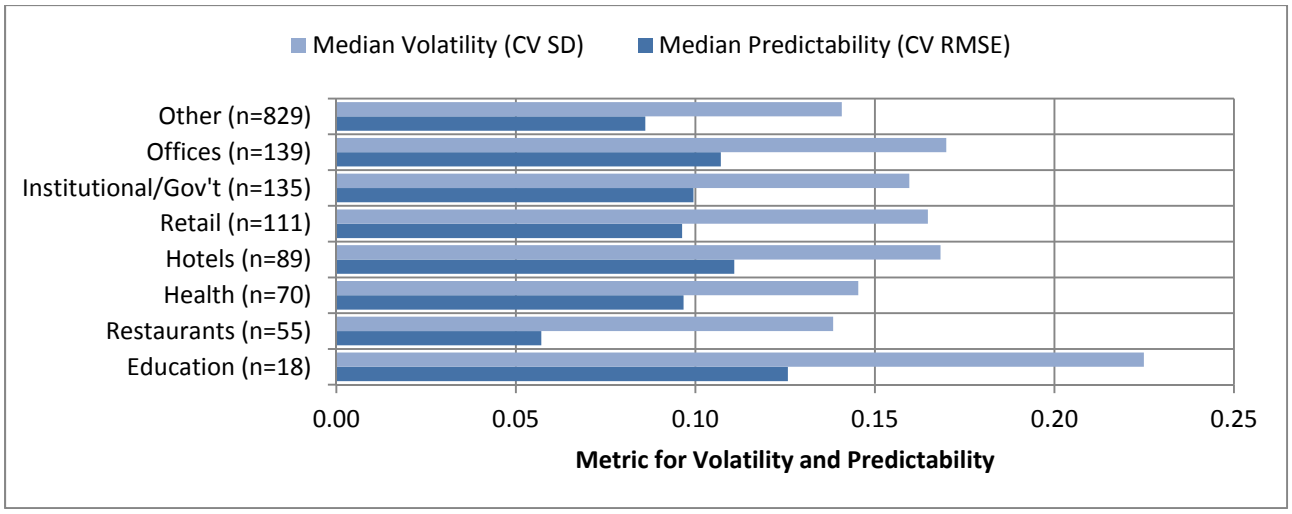



Figure 5. Variability and predictability for various customers segments.

Customers that are highly weather sensitive, as measured by the correlation between monthly consumption and cooling and heating degree days, have the highest variability but are also more predictable than less weather sensitive customers. We know why energy consumption for these customers varies; we know less about what drives variation in consumption for less weather sensitive customers.

The effect of weather is also observed in the variability across geographic areas in PG&E's territory. Customers in hotter inland areas such as Kern and Fresno, where temperature often exceed 105°F, tend to be the most variable, but much of this variability is explained by weather. Patterns regarding which industries are more or less predictable are unclear, except for two industries: educational facilities tend to have more variability and are harder to predict; while restaurants are among the most predictable industries.

However, the volatility of a customer's month-to-month consumption was the best indicator of predictability. This is not surprising since variability is essentially an upper bound for predictability. While the above information was useful to help identify potential screens to avoid hard-to-predict customers, we note that screens based on weather sensitivity, business type, location and historical variability are not very useful for identifying which customers are likely to experience disruptive changes leading to a large change in year-to-year consumption.

## Accuracy of Energy Saving Estimates With and Without Pre-Screening

Before summarizing the accuracy of the results with and without pre-screening, it is important to highlight the distinction between baseline errors, impact errors (energy savings), and payment errors. Analyzing individual sites produces baseline errors that are relatively large. Even when a baseline is relatively accurate for an individual customer, it can still produce large errors in estimated savings because the ability to accurately detect energy savings is tied to the size of the signal. Detecting small energy savings, such as savings less the 10%, from unexplained variation is extremely difficult for specific customers. The structure of payments affects the extent by which energy saving errors translate into payment error. For example, some programs pay customers incentives or rebates when the baseline error favors the customer, but do not charge the customer if baseline error does not favor the customer (e.g., the baseline indicates an increase in energy consumption).

Figure 6 illustrates the distinction between baseline, impact (energy savings) and payment errors for the same set of customers. Baseline errors are not tied to the magnitude of true energy savings, while impact and payments errors are magnified when true energy savings are smaller. To illustrate the distinction between baseline and impact error, consider an instance where the correct energy savings is 5 % percent but the baseline overestimates what energy consumption would have been by 5%. Such a baseline would incorrectly estimate energy savings of 10% - double the true savings. Conversely, if the baseline had underestimated by 5% it would negate the true energy savings.

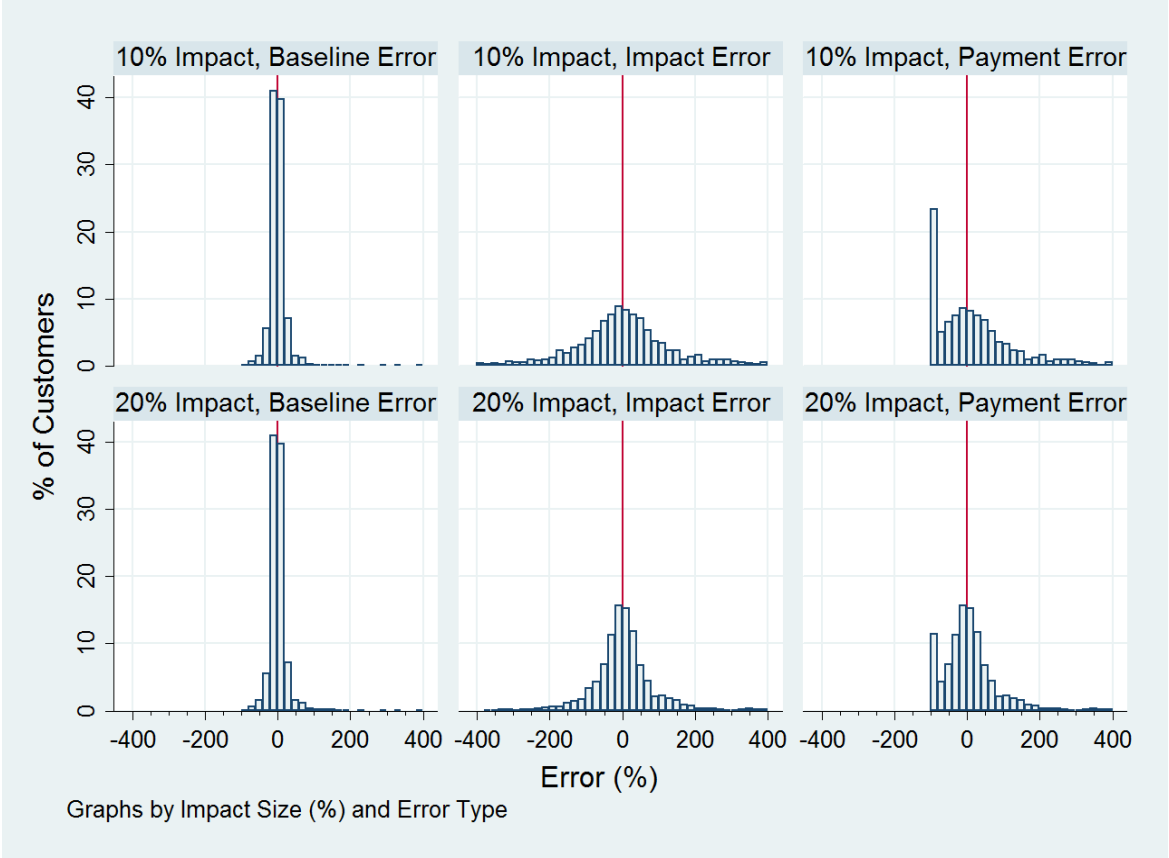



Figure 6. Distribution of baseline, energy savings and payment error

Figure 7 summarizes the distribution of baseline errors by industry and location. The baselines were estimated using more granular smart meter data and controlled for the effect of weather, day of week and seasonality on consumption. They were developed using the first year of data and used to predict what consumption would have been absent energy efficiency for the second year. This allows an assessment of accuracy by comparing the baseline estimates against the correct counterfactual. The most notable observation is that even after modeling energy consumption patterns with smart meter data, individual customer estimates still have a large amount of error irrespective of industry or location.

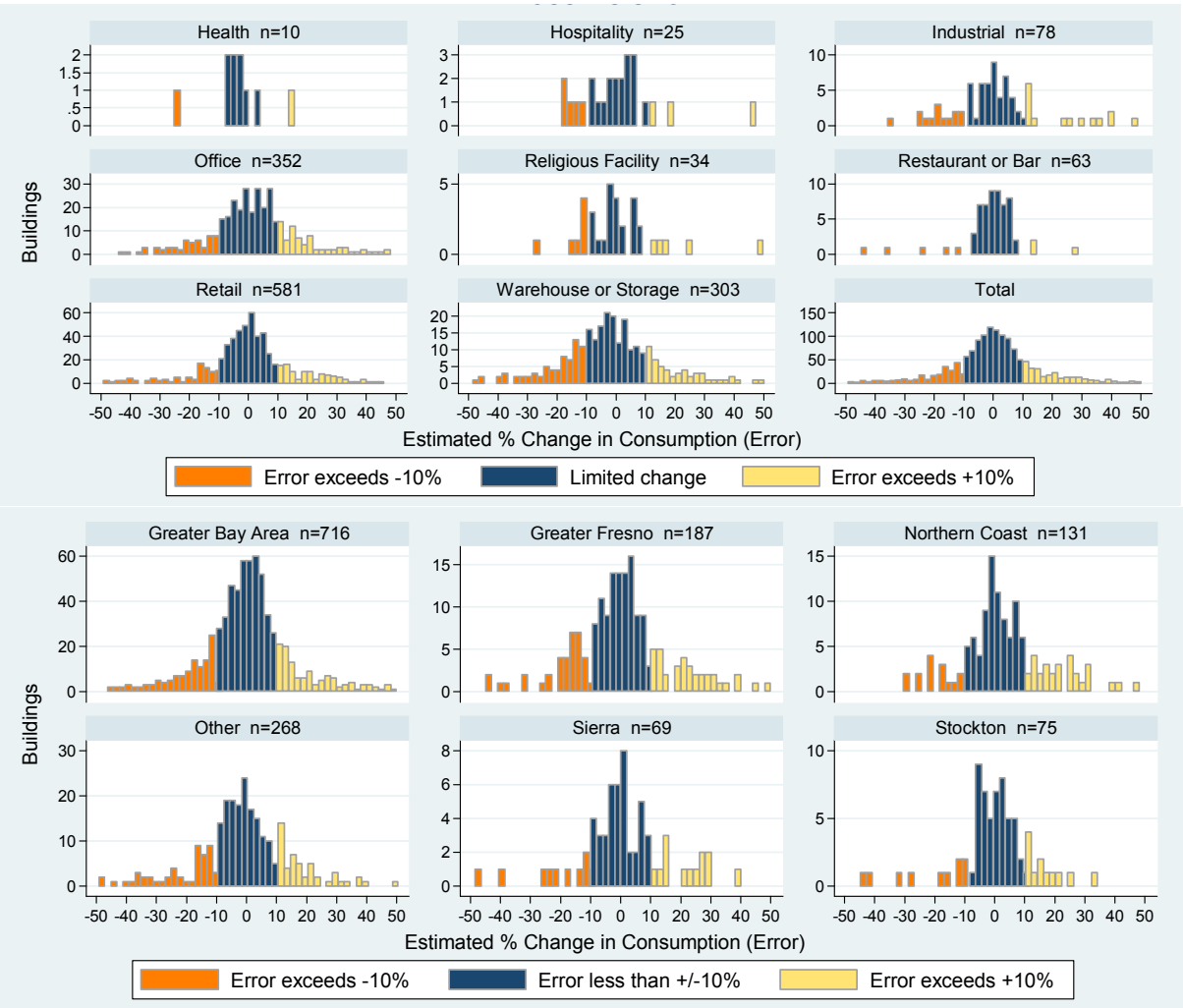



Figure 7. Distribution of baseline error by industry and geographic area.

Figure 8 summarizes the effect of applying various screens to identify and exclude hard-to-predict customers and compares the range of baselines with screening against the range without pre-screening. For brevity, we summarized 16 of the screening criteria tested. In total, 80% of eligible customers fall within the bars. In all cases the baseline error for the median customers is near zero. However, the baseline errors for a large share of customers is as large if not larger than realistic energy savings that can be attained with whole building retrofits. No metric or screen is perfect – the best reduce the range of baseline errors in half.

The application of screening comes at a cost. Imposing screening criteria also leads to exclusion of a significant share of customers, many of whom do not experience substantial changes in year-to-year consumption. Figure 9 summarizes this tradeoff. It depicts how two types of errors – excluding predictable customers and including hard-to-predict customers – vary for different screening criteria.

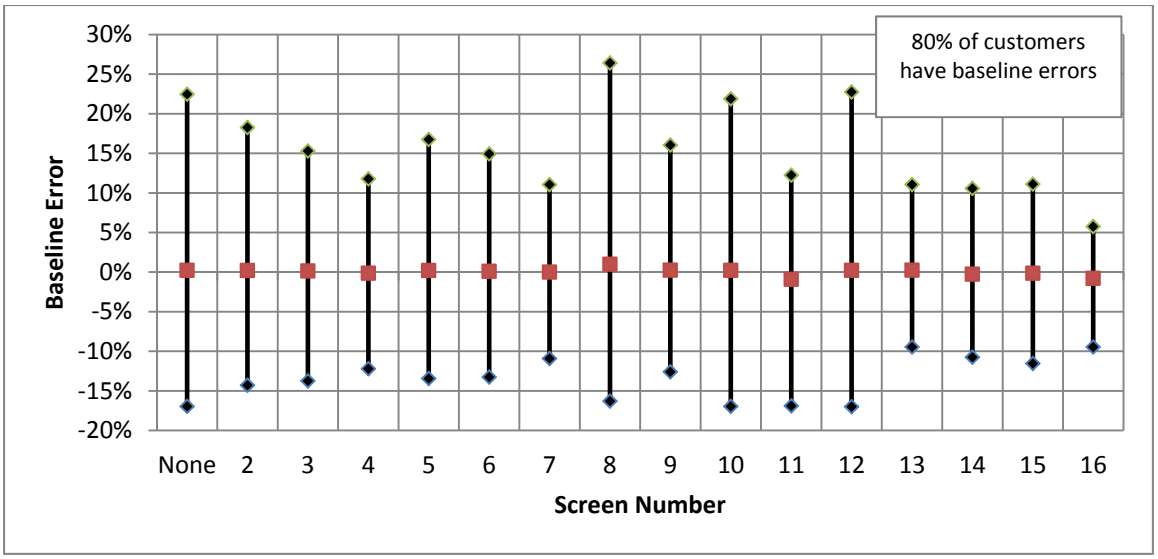



Figure 8. Comparison of baseline errors with and without pre-screening.

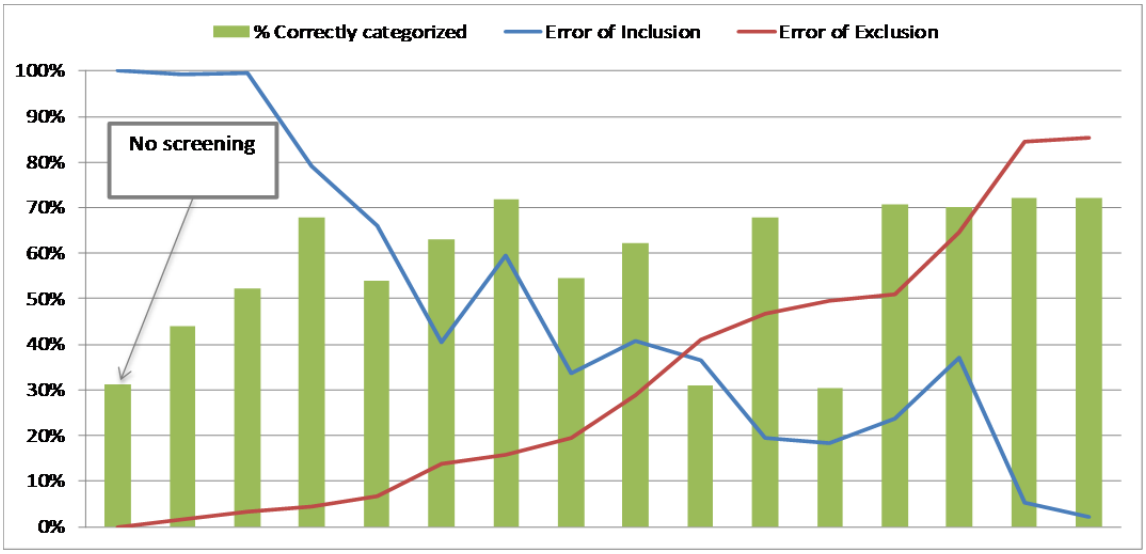



Figure 9. Errors of inclusion and exclusion due to pre-screening.

Use of customer specific estimates of energy savings requires extreme caution, whether for impact evaluation, program design, or implementation. While time series regression analysis make use of more granular data and can better control for the effects of weather on energy consumption, the largest source of year-to-year changes in energy use is not weather, but changes building operations and occupant behavior that often cannot be observed by evaluators.

It is possible that results could be improved marginally through better modeling, additional data, or better screens. But given the extensive testing conducted as part of this study any improvements are likely to be small and unable to address the fundamental flaws of estimating customers specific energy savings based on usage patterns prior to energy efficiency upgrades. Perhaps the best potential for improving accuracy of energy savings measurement may be obtained by:

- Pooling or aggregating across customers;
- Incorporating high quality control groups to better establish behavior absent energy efficiency improvements; and
- Limiting applications of customer specific savings to customers expected to provide deep percent reductions in consumption.

Figure 10 illustrates the power of pooling across customers. The customer specific estimates were aggregated assuming different sample sizes. We repeated the aggregation two hundred times for each sample size using different samples (a process known as bootstrapping) to understand the range of baseline errors when results are aggregated. While the graph is illustrative, it does not fully factor in that customers who self-select into a program are not necessarily random. It also does not apply to approaches that also make use of external control groups in addition to participant data before and after energy efficiency upgrades.

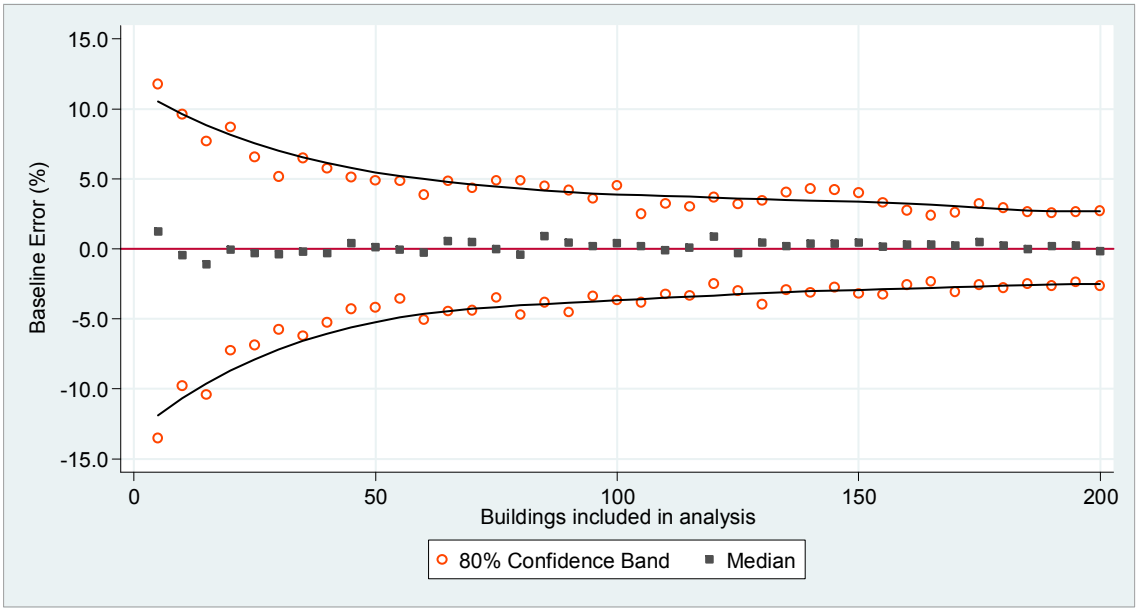



Figure 10. Baselines errors with aggregation.

## Key Findings and Conclusions

Our main conclusions are:

1. A substantial share of customers experience year-to-year change in energy consumption that is unrelated to weather patterns or energy efficiency. In PG&E's territory, roughly 40% of buildings experience year to year changes in electricity consumption in excess of 10% (based on the random sample).
2. The best indicator of predictable energy consumption patterns was a customer's volatility in month-to-month consumption. Other factors such as weather sensitivity, location, and business type were also useful indicator of predictability. However, a large share of

customers experienced disruptive changes in energy use that could not be explained by year-to-year variation in weather, weather sensitivity, location, or business type.

3. The accuracy of customer specific energy savings estimates produced using time series regression models can vary considerably for individual sites. In general, analyses that rely on whole building usage patterns perform better when aggregated across multiple customers.
4. Pre-screening can help identify hard-to-predict customers. However, no single quantitative screen appears to be effective at filtering out all hard-to-predict customers, and each screen also excludes a large number of customers who are indeed predictable. Qualitative screens may prove helpful, but these were out of scope for this analysis.
5. The best quantitative screens identified reduced the range of baseline errors by half. While this is a significant improvement, even after screening, the baseline error for many customers was as large if not larger than realistic whole building energy savings from energy efficiency improvements. As a result, it is important to limit applications of customer specific savings to customers expected to deliver deep percent reductions in whole building energy use.

Smart meter data has significant potential to improve evaluations through the use of external control groups and large samples. It also can significantly improve targeting of high potential customers. However, to date, much of the attention and effort has been focused on using smart meter data to produce customer specific energy savings based on time series of energy use patterns before and after energy efficiency improvements. As shown in the paper, this emphasis has limitations; and caution needs to be exercised in using customer specific savings estimates. We recommend shifting the emphasis of research regarding applications of smart meter data to investigating the implications of larger sample sizes and matched control groups on the accuracy of energy savings estimates.

Analyses with whole building data may prove to be more accurate than other approaches, such as short and mid-term end-use metering. The study did not compare the accuracy of different methods but rather focused exclusively on whether time series regressions could accurately produce customer-specific energy savings. Alternate evaluation methods have historical reasons for their use and advantages and disadvantages. For example, short- and mid-term end use metering suffer from many of the same shortcomings as analysis using whole building data. They rely exclusively on data regarding electricity use before and after installation of energy efficient equipment without an external control group. While the end-use energy savings (the signal) is larger and the background noise is smaller with end-use data, the sample sizes may be far smaller and the data collection is often relatively short (a month before and after installation), requiring extrapolation of results far out of sample.

There are several areas for additional research. One possibility is to explore if and how customer specific estimates of energy savings could be improved. Studies could explore different regression model specifications, incorporate additional data such as occupancy (provided the data is consistently available to evaluators rather than provided only when it suits participants), or explore different analytic techniques. In our view, the largest potential benefit is in comparing different methods for program level evaluation and the use of smart meter data to identify high potential opportunities. How do larger sample sizes, made possible by smart meter data, affect precision? Does this increase in precision outweigh the additional noise inherent in moving from end-use metering to whole building data? To what degree does the use of external control groups developed by matching of pre-enrollment load patterns improve accuracy? Does

using smart meter data for evaluation increase precision and lower costs? Does use of smart meter data allow utilities to better identify customers who have higher than typical energy use intensity after accounting for their business type, hours of operations (which often can be inferred from smart meter data), and building square footage?

## References


*George, S., Schellengberg, J, Holmberg, S. 2012. Impact Evaluation of Southern California Edison's 10 for 10 Rebate Program.* In proceedings of *2013 International Energy Program Evaluation Conference*. August 13-15 , 2013, Chicago, IL.

Price, P., Jump, D, Granderson, J. Sohn, M. Addy, N. 2013. *Commercial Building Energy Baseline Modeling Software: Performance Metrics and Method Testing with Open Source Models and Implications for Proprietary Software Testing.* Available at: http://www.etcc-ca.com/sites/default/files/reports/ET12PGE5312_EMIS_SoftwareBaselineModeling_ModelAnalysis_0.pdf

Shadish, W, Cook T., Campbell D. 2002. "Experimental and Quasi-experimental Designs for Causal Inference." (Boston: Houghton Mifflin Company, 2002).