A Smart Approach to Analyzing Smart Meter Data

Ted Helvoigt, Evergreen Economics (Lead Author)
Steve Grover, Evergreen Economics
John Cornwell, Evergreen Economics
Sarah Monohon, Evergreen Economics

ABSTRACT

The wealth of information contained within advanced metering infrastructure (AMI) data offers great promise to utilities in designing and understanding the impacts of energy efficiency and demand-side management programs. At the same time, fully capturing the information contained within AMI data is challenging due to the sheer volume of data.

We discuss some of the methods we employed for a recently completed research project for the California Investor-Owned Utilities in which we employed a random coefficient model to estimate more than 1,000 unique load shapes, each representing one of 20 different home-bins on one of more than 50 different combinations of cooling degree-days (CDD) and heating degree-days (HDD). Unlike standard methods of regression analysis, which fit a single line through a scatter of data, the random coefficient model fits a unique regression line to each load shape while simultaneously accounting for correlations in energy use across all load shapes.

This paper will be of interest to any researcher working with AMI data to understand time-of-day energy use at the building level.

We begin this paper by discussing the abundance of data generated by AMI—are all these data too much of a good thing? We then briefly discuss the fixed-effects model for estimating a billing regression and then present the random coefficient model. We conclude with a discussion of potential applications for the random coefficient model with respect to AMI data.

AMI Data…Too Much of a Good Thing?

For decades, evaluators specified billing regression models to address only those questions that were answerable with the data at hand. Since utilities typically aggregate a customer’s continuous electricity use data in monthly billing cycles, the evaluator typically has 12 observations of electricity use per customer per year. With these data, the evaluator can develop models of monthly energy use and, if the purpose of the billing regression is to estimate the impacts of an energy efficiency program, will be able to produce estimates of monthly and annual energy use or savings. Estimates of monthly energy use or savings may satisfy the needs of the client for many evaluation projects; for these projects, interval metered data measured in 15-minute, 30-minute, or one-hour intervals provide no additional value but may impose a cost on the evaluator due to their sheer volume. For interval data to be of value for evaluation, the question(s) addressed by the evaluator must be deeper than “what were the annual savings?”

Most evaluators work with software packages that have no (theoretical) limit on the number of observations in a data file. Unfortunately, computers do have processing and storage limits, and very large files will slow down all steps of the data analysis and modeling. To put this into the context of an impact evaluation, let us examine a scenario in which you are evaluating a residential energy efficiency program with 10,000 participants. The client provides you with two years of pre-program data and one year of post-program data for each of the 10,000 participants and for 5,000 non-participants. Assuming no data are missing, if the client provides you with monthly billing data, you will receive a file with 540,000 records.[1] If instead the client provides you with interval data measured on an hourly basis, you will receive a file with 394,200,000 records.[2]

The file of hourly interval data contains 730 times as many records as the file of monthly billing data. If the client only cares about the annual and/or seasonal savings from the energy efficiency program, the rational evaluator would first aggregate the hourly file to a monthly level and then estimate the billing regression. If, however, the client needs to know how the energy efficiency program affects energy use at different times of the day—e.g. during the afternoon peak—then the evaluator can only address this question using hourly data.

Figure 1 shows an example of one month of electricity consumption data measured on an hourly basis for a single home in California during June 2012. The jagged blue line shows actual electricity use for each hour, while the red line simply shows the average hourly electricity use during the month. The hourly consumption data provides highly detailed information on how this household used electricity each day throughout the month. Comparatively, the average hourly electricity usage reveals no more about the electricity use of the customer than does a monthly billing record. AMI provides potentially important information about how a customer uses electricity through the day, the effect that temperature has on daily electricity use, and (potentially) the impact that participation in an energy efficiency program has on energy use during periods of peak demand.

While data on average monthly electricity use may be sufficient to develop estimates of electricity savings, it is likely that regulators and policy makers will demand that energy efficiency programs not only demonstrate that they save energy, but also reveal in which hours of the day, in which days of the week, and in which weeks of the year those savings occurred. While most utilities and other program administrators may only be concerned with estimates of annual electricity savings from an energy efficiency program, this is largely an artifact of the data historically available for evaluation—monthly billing data. This will change as AMI becomes more widespread, as those in the industry increasingly understand how data from AMI can inform program design, and perhaps most importantly, as the measures and programs that utilities rely on to achieve energy savings become more complex.

[1] (10,000 Parts + 5,000 Non-parts) * (36 Pre- and Post-months) = 540,000 observations.

[2] (10,000 Parts + 5,000 Non-parts) * (365 days per year * 3 years) * 24 hours per day = 394,200,000 observations.

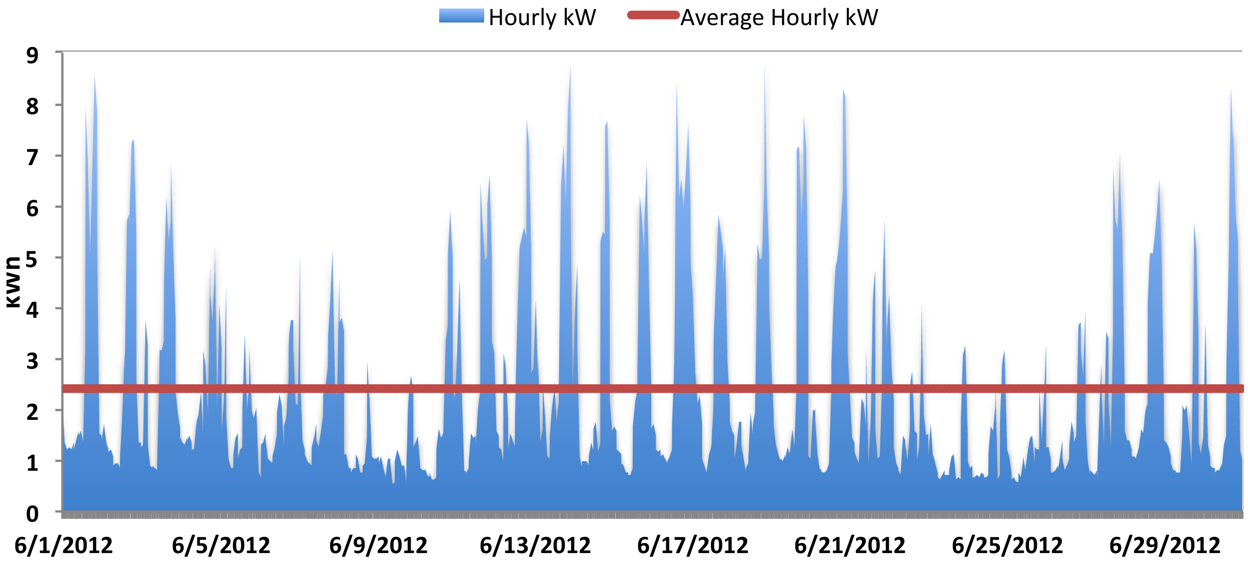

Figure 1. Hourly kWh usage for one home in Southern California. *Source:* Analysis by Evergreen Economics of data provided by Southern California Edison.

An assumption made either explicitly or implicitly in most residential billing analyses is that the value of each estimated coefficient is constant across the homes in the analysis. While this assumption simplifies the analysis, allowing the evaluator to estimate the regression using a fixed-effects or other standard modeling approach, it is unrealistic to assume the coefficients do not vary across customers.[3] It is more likely that participants of an energy efficiency program will experience a distribution of energy use responses, with some customers experiencing little or no energy savings, some experiencing substantial energy savings, and the remainder falling somewhere in the middle. When considering customers' responses to an energy efficiency program in this way, it is reasonable to view the coefficients from a billing regression as random variables, with each customer having their own set of regression coefficients representing their particular response to the program.

Fixed-Effects Regression Model

The fixed-effects regression model is a standard approach for estimating a billing regression using data on a panel of residential customers.[4] The popularity of the fixed-effects regression model is due to it being easy to implement, included in most statistical software packages, relatively easy to explain to a non-technical audience, and, most importantly, consistent with the typical assumption that the values of the explanatory variables are non-random. The fixed-effects regression model is also a popular choice because it controls for unobserved heterogeneity among observational units (e.g. residential customers). Empirically,

[3] It is important to note that the evaluator may believe the coefficients differ across homes, but does not believe the distribution of these differences to be of practical importance for the evaluation.

[4] For information on the fixed-effects regression model, see Greene 2003 or another intermediate econometric textbook. Evaluators do also use other modeling approaches such as ordinary least squares (OLS) and random effects (Greene 2003); however, it is sufficient for our purposes to consider only the fixed-effects model as the standard choice for evaluators to conduct billing regression.

there are several ways to accomplish this, including adding a dummy variable for each observational unit. The result is that the model estimates a unique y-intercept (i.e., a fixed effect) for each observational unit and a single vector of slope coefficients that represents the mean response across all observational units. The shape of the estimated regression line, therefore, is the same for all observational units.

Figure 2 shows an example of hourly electricity use for a sample of homes in Southern California that participated in Southern California Edison’s (SCE's) Quality Installation (QI) program and an estimated regression line that represents the mean hourly electricity use for each hour of the day. Because the fixed-effects regression model is an extension of the ordinary least squares (OLS) model, the criterion for fitting the regression is to minimize the sum of squared errors between the regression line and the data point. In doing so, the regression line is not fit to the individual electricity load shapes shown in Figure 2 but rather is fit to the individual points.

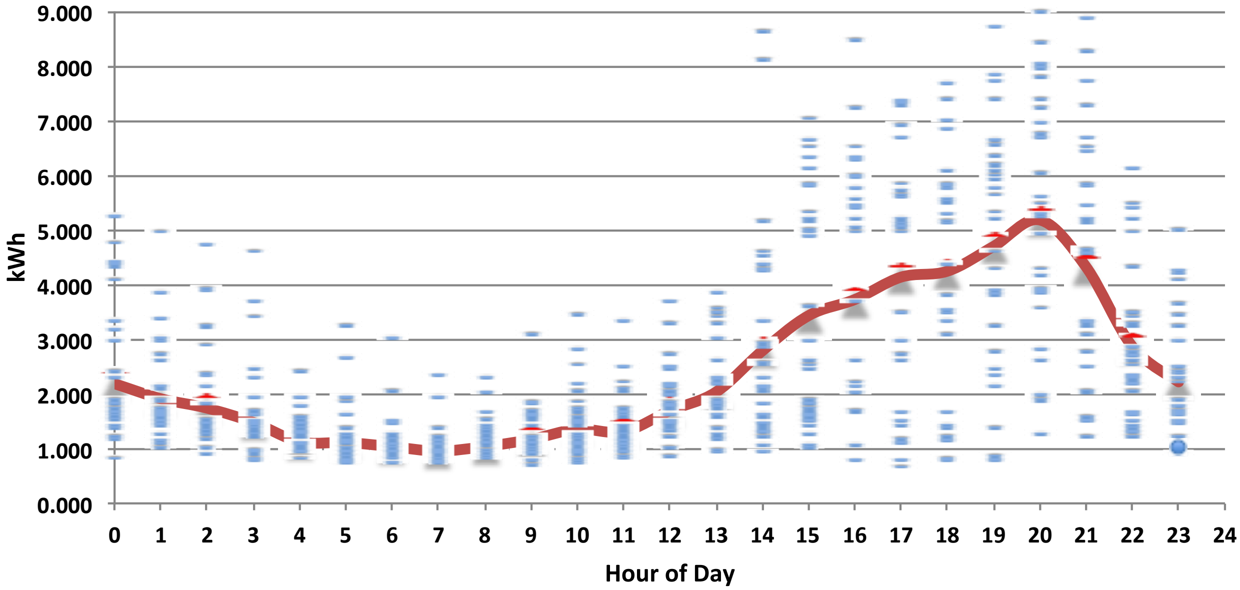

Figure 2. Mean estimated load shape for one day of hourly electricity use for a sample of QI participants, based on a fixed-effects modeling approach. *Source*: Analysis by Evergreen Economics of data provided by SCE.

Random Coefficient Model

While the fixed-effects regression model focuses on the average response and how it changes as values of an explanatory variable change, an alternative is to instead first focus on the trajectory of each observational unit.[5] This alternative approach, the random coefficient model, explicitly considers the trajectory of each observational unit—e.g. the unique response of electricity use to time, temperature, or other explanatory variable. Whereas the fixed-effects model estimates a unique y-intercept for each observational unit, but a single vector of slope coefficients for all observational units, the random coefficient model estimates a unique y-intercept and vector of slope coefficients for each observational unit.

[5] The observational unit could be an individual residential or non-residential customer, or some aggregation of customers based on similar historical behavior, geography, temperature, or other criteria.

What this means for billing analysis is that the random coefficient approach allows the evaluator to model the behavior of an individual observational unit, which could be a single residence or business, or, as we describe below, aggregate data for a group of residences with similar historical energy consumption. For example, Figure 3 shows the electricity-use trajectories (load shapes) for one day for four of the residential customers who are participants in the QI program shown earlier in Figure 2.

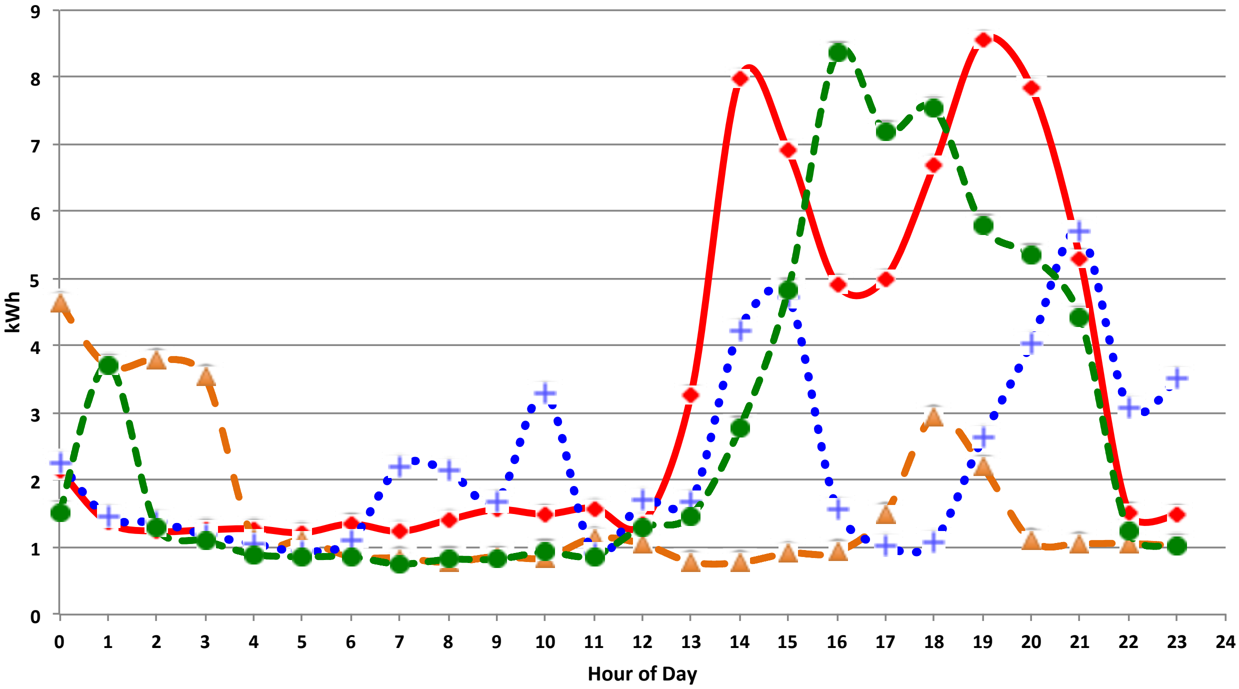

Figure 3. One day of hourly electricity use by four participants of QI program. *Source*: Analysis by Evergreen Economics of data provided by SCE.

If the evaluator is interested in modeling the load shape of each of these customers, he or she could do this by estimating a separate regression model for each customer. The evaluator could also model the average behavior of the four customers using a fixed-effects regression model. With a small number of customers in the program, this requires little effort; however, if there are hundreds or thousands of program participants, then estimating separate models, while technically feasible, will be extremely resource intensive.

Alternatively, the evaluator could estimate a random coefficient model, which is a type of multilevel regression model. For our example, the first-level model consists of individual equations estimated for each of the four participants in Figure 4. Because there is likely at least some degree of temporal correlation between residuals, the equations are estimated using generalized least squares (GLS).[6] When heteroskedasticity and/or autocorrelation are present in the data, the GLS estimator is asymptotically more efficient than OLS. Equation 1 shows the first

[6] Generalized least squares is a method for estimating the unknown parameters in a linear regression model when there is correlation between the residuals, a violation of the classical regression model (see Greene 2003 or other econometric text).

level model of the Hildreth and Houck (1968) random coefficient model, which we applied in a recently completed California statewide research project.

Equation 1. First level *individual* model

$$kWh_{it} = \beta_{0i} + \beta_{1i}CDH_{it} + \beta_{ki}X_{ik} + \varepsilon_{ij}$$

$$\begin{aligned}
&\mathit{Where}: \\
kWh_{it} &= \mathit{Electricity\ usage\ for\ the\ i\text{-}th\ home\ in\ the\ t\text{-}th\ (hourly)\ interval} \\
CDH_{it} &= \mathit{Cooling\ degree\ hours\ for\ the\ i\text{-}th\ home\ in\ the\ t\text{-}th\ (hourly)\ interval} \\
X_{ki} &= \mathit{Array\ of\ k\ hourly\ indicators\ for\ i\text{-}th\ home} \\
\beta_{0i}, \beta_{1i}, \beta_{ki} &= \mathit{Parameters\ to\ be\ estimated\ for\ the\ i\text{-}th\ home} \\
\varepsilon_{it} &= \mathit{Random\ error\ term\ for\ i\text{-}th\ home\ in\ the\ t\text{-}th\ (hourly)\ interval}
\end{aligned}$$

The intercept and the slope coefficients for the second-level model (also referred to as the population model) is the mean of the intercepts and the slope coefficients estimated from the first-level model.

Equation 2. Second level *population* model[7]

$$\mathrm{E}\left[\beta_{0}\right] = n^{-1}\sum_{i=1}^{n}\widehat{\beta}_{0i} \qquad \mathrm{E}\left[\beta_{1}\right] = n^{-1}\sum_{i=1}^{n}\widehat{\beta}_{1i} \qquad \mathrm{E}\left[\beta_{k}\right] = n^{-1}\sum_{i=1}^{n}\widehat{\beta}_{ki}$$

$$\begin{aligned}
&\mathit{Where}: \\
\beta_{0}, \beta_{1}, \beta_{k} &= \mathit{Population\ parameters} \\
\widehat{\beta}_{0i}, \widehat{\beta}_{1i}, \widehat{\beta}_{ki} &= \mathit{Estimated\ parameters\ from\ the\ individual\ regressions}
\end{aligned}$$

The second-level model represents the average of the estimated trajectories for each first-level model. Thus, even if the focus of the evaluator's interest is on the individual observational units, the random coefficient model still estimates a population-based model with intercept and slope coefficients drawn from the multivariate probability distribution of intercept and slope coefficients estimated for all individual observational units included in the model.

It follows from Equation 2 that we can represent the coefficients from the individual regressions as random deviation from the population mean.

$$\widehat{\beta}_{0i} = \mathrm{E}\left[\beta_{0}\right] + b_{0i} \qquad \widehat{\beta}_{1i} = \mathrm{E}\left[\beta_{1}\right] + b_{1i} \qquad \widehat{\beta}_{ki} = \mathrm{E}\left[\beta_{k}\right] + b_{ki}$$

Each individual b_i represents a random effect describing how the intercept and slope of an individual deviates from the population mean (Davidian 2005). The b_i vectors have a mean of zero and a covariance matrix that describes the variation (Davidian 2005).

[7] Equation 2 assumes equal weight for each individual. In practice, the second-level model is often computed as the weighted average of the individual coefficients. Model estimation conducted in Limdep econometric software (Greene 2002).

The random coefficient model explicitly accounts for two separate sources of variability commonly found in interval energy-use data.

Within-Subject Variability

In the first-level model, the random coefficient model accounts for within-subject variability, which, for AMI data, might be variation in energy usage through the day by an individual residential or non-residential customer (or aggregation of customers). This variability may be due to multiple factors, including changes in temperature, idiosyncratic actions by the resident(s) or business, and even measurement error.

Among-Subject Variability

In the second-level model, the random coefficient model accounts for among-subject variability, which for a model based on AMI data might be variation in energy use across customers and temperature. In this way, the random coefficient model accounts for variability in electricity use between homes, through time, and across varying temperatures.

Consider for example a modeling exercise in which the observational unit is defined by household and by average daily outside temperature, and the variable of interest is hourly electricity usage (this is essentially the approach we describe below). In this exercise, the evaluator is interested in the trajectory of daily electricity load shapes by household and temperature. Figure 4 shows an example of one type of among-subject variability. Each of the three lines in the figure represents a daily electricity load shape for the same home, but for three different average daily temperatures (70°F, 79°F, and 88°F).[8]

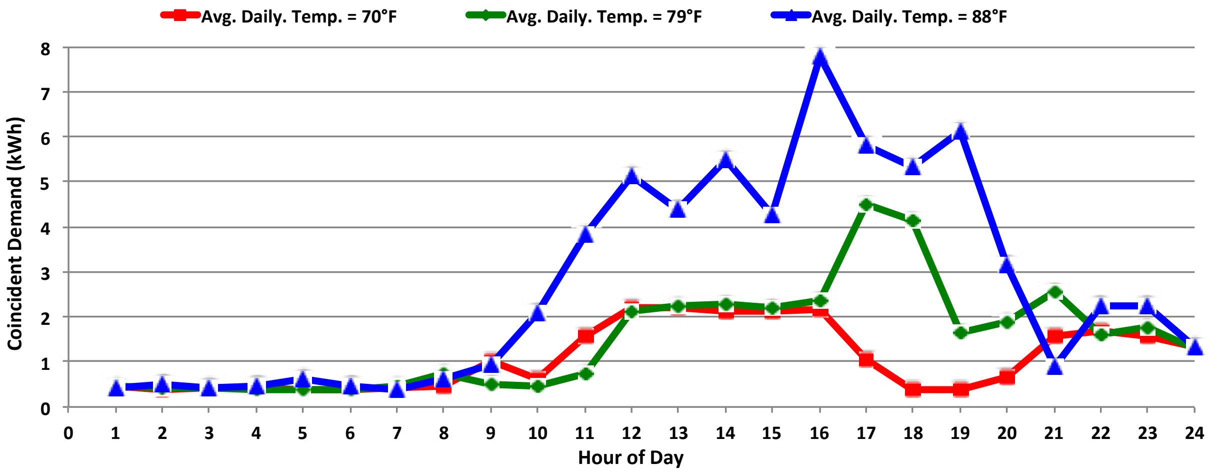

Figure 4. Hourly electricity use for same home on three different days, summer 2013. *Source*: Analysis by Evergreen Economics of data provided by SCE.

Figure 5 shows an example of another type of among-subject variability. Each of the three lines in the figure represents a daily electricity load shape for a different home experiencing

[8] Based on a base temperature of 65°F, these average daily temperatures would correspond with a CDD of 5, 14, and 23, respectively.

the same average daily temperature (75°F).[9] For each hour of the day, electricity use varied among the three homes, with the variability being greatest after the noon hour (and at 1 a.m.).

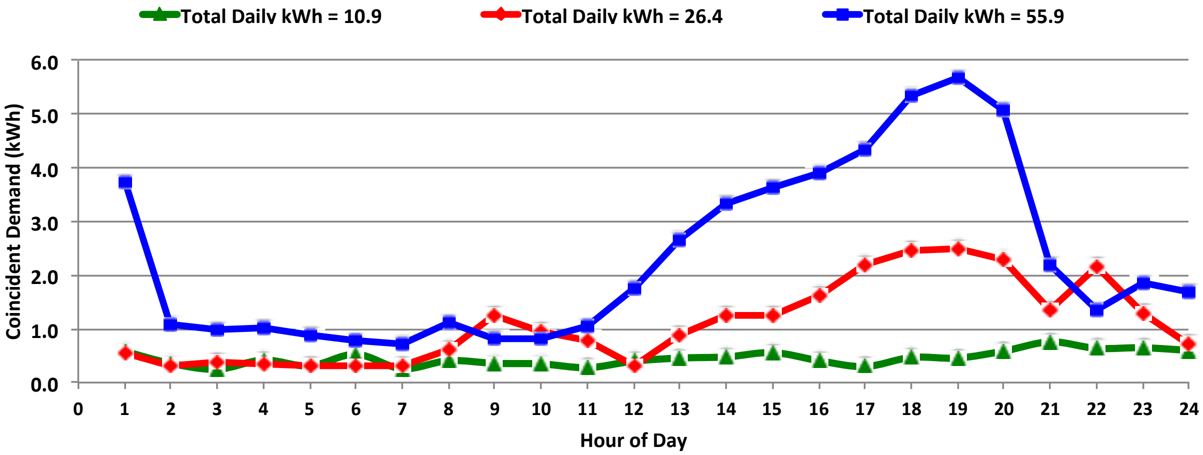

Figure 5. Three different homes experiencing the same average daily temperature of 75 degrees Fahrenheit (10 CDD). *Source*: Analysis by Evergreen Economics of data provided by SCE.

Potential Applications of the Random Coefficient Model

We believe the random coefficient model represents a significant and positive departure from current approaches to analyzing AMI data and estimating impacts from energy efficiency and demand response programs. While analytically and conceptually more sophisticated than the fixed-effects model, the random coefficient model takes greater advantage of the information provided by AMI data. As utilities continue to migrate their customers to interval meters, we believe it is natural that they embrace methods of analysis that more fully exploit the abundant information contained in AMI data. The random coefficient model we describe in this report does this.

Evaluators can use the random coefficient model to develop estimates of impacts for energy efficiency and demand response programs that are specific to individual customers and outdoor temperatures. There are three areas of utility operations in which we believe application of the random coefficient modeling approach may be beneficial.

1. Evaluation of energy efficiency and demand response programs;
2. Energy efficiency and demand response program planning and targeting; and
3. Short-term and long-term load forecasting.

[9] CDD of 10 based on a base temperature of 65°F.

1. Evaluation of Energy Efficiency and Demand Response Programs

By employing a hierarchical modeling approach such as the random coefficient model, an evaluator can efficiently develop portfolios of daily load shapes for individual homes or groups of similar homes tailored to specific daily temperatures and days of the week. With the information produced from the random coefficient model, the evaluator can quantify how energy consumption behavior varies across groups of homes, as well as across the range of observed daily temperatures and days of the week. And, because the random coefficient model explicitly considers the unique effect that temperature and day of the week have on each home or group of homes, the evaluator can develop precise forecasts of energy consumption for each participant in the energy efficiency program.

Comparing the portfolio of daily load shapes in the pre- and post-periods allows the evaluator to identify where energy savings are occurring for each home, on each day of the week, and at each temperature. For example, in our recently completed analysis, we found significant variability in energy savings among the 20 pre-defined cohorts of participants of SCE’s QI program, ranging from no energy savings for some cohorts to significant savings for others. This is actionable information the utility could use to target those customers with the greatest expected energy savings for future participation in the program.

We were also able to pinpoint the days and hours in which the greatest energy savings occurred for the QI program—hot summer days in the mid-to-late afternoon. This finding was not necessarily surprising given the nature of the HVAC program. Nevertheless, without AMI data, evaluators could only conjecture on what days and in which hours savings occurred. Applying the hierarchical random coefficient model allowed us to accurately quantify those savings by hourly temperature for each time of day for each day of the week for each cohort of participants.

The random coefficient model is also a valuable tool for demonstrating attribution of savings to an energy efficiency program. Because the model provides disaggregated estimates of energy savings by time of day, season of year, and a customer’s historical energy usage, an evaluator can compare model results to the ex ante expectations of the program to determine if the pattern of energy savings occurred in a way that was consistent with the design of the program. For example, in the QI HVAC program, our ex ante expectation was that most energy savings would occur on hot days during the peak period, and would be especially great for those customers with historically high energy usage. The results of our analysis corroborated these expectations, suggesting that the savings we observed were indeed attributable to the QI HVAC program.

2. Energy Efficiency and Demand Response Program Planning and Targeting

The random coefficient model is a valuable tool for energy efficiency program planning and targeting, which program planners can use to improve the effectiveness of their programs in several ways:

1. Increase the cost effectiveness and efficiency of program marketing and participation criteria;
2. Improve the equity of energy efficiency programs by providing potential participants with realistic estimates of the energy and money they will save;

3. Develop “smart” offers to customers that accurately inform them of the potential savings from one or more energy efficiency programs based on their unique characteristics; and
4. More optimally develop energy efficiency programs based on the performance of current and former participants and the characteristics of potential participants.

By segmenting homes based on energy consumption and non-energy attributes (e.g. home size, location, etc.), program planners can identify which homes save the most energy from installation of an energy efficiency measure. Identification of these high savings homes will allow program planners to more effectively target customers, or adjust measure and program offerings to maximize energy savings.

Program planners can use the random coefficient model to develop customer-specific predictions of likely energy savings associated with various energy efficiency programs, thus empowering the utility to target the most beneficial programs to each customer. The random coefficient model would allow the utility to project how much energy a customer will save and during what time of year and day those savings will occur. With this information, the utility can estimate customer-specific cost effectiveness of participating in the program.

3. Short-Term and Long-Term Load Forecasting

The attributes of the random coefficient model that make it valuable for program evaluation, design, and targeting, also make it valuable for load forecasting. With the multiple levels of estimated coefficients from a random coefficient model, a utility could develop forecasts of energy use for residential or non-residential customers that explicitly account for thousands of different usage groups (based on historical energy usage, installed equipment, or other attributes) experiencing a range of alternative weather conditions. Such an approach would allow the utility to develop realistic projections of energy demand under extreme weather conditions or for events that can have a large impact on the grid, over a long (multi-year) planning horizon, as well as over a very short (hourly or daily) horizon.

Discussion

As utilities continue to invest in AMI technology, evaluators and others analyzing AMI data are faced with two significant challenges: (1) developing meaningful and actionable results that take advantage of the wealth of information contained within the AMI data, and (2) obtaining these results quickly and efficiently. We believe the random coefficient model meets both of these challenges. The hierarchical structure of the model explicitly accounts for the variability in energy-consuming behavior among utility customers. Rather than simply model the average response to an energy efficiency program, the evaluator is able to model the response of individual customers or customer segments. The hierarchical structure is also important from a pragmatic standpoint—the evaluator can obtain these customer-level results (i.e., vector of regression coefficients) quickly. Finally, by estimating these first-level models using GLS, the model accounts for the likely, but unknown, correlation between residuals.

We are aware of at least one study that applies a random coefficient modeling approach to estimate electricity load shapes (Fiebig, Bartels, and Aigner 1991). There may be others. The focus of the authors' analysis is the development of end-use load profiles using conditional demand analysis. The random coefficient model is used to estimate 24 hour-specific regression models to obtain estimates of electricity use by end-use.

References

Davidian, M. "ST 732: Applied Longitudinal Data Analysis." Chapter 9. Lecture, Department of Statistics, North Carolina State University, Raleigh, NC. 2005. http://down.cenet.org.cn/upfile/79/200711181188195.pdf

Fiebig, D.G., R. Bartels, and D.J. Aigner. 1991. "A Random Coefficient Approach to the Estimation of Residential End-use Load Profiles." *Journal of Econometrics* 50 (1991): 291-327.

Greene, W.H. 2002. *Limdep Version 8.0 Reference Guide*. New York: Econometric Software, Inc.

Greene, W.H. 2003. *Econometric Analysis, 6th Edition*. Englewood Cliffs, NJ: Prentice Hall.

Hildreth, C. and C. Houck. 1968. "Some Estimators for a Linear Model with Random Coefficients." *Journal of the American Statistical Association* 63: 584-95.