Is More Always Better? A Comparison of Billing Regression Results Using Monthly, Daily and Hourly AMI Data

John Cornwell, Evergreen Economics, Portland, OR
Stephen Grover, Evergreen Economics, Portland, OR

ABSTRACT

The evaluation community has long known that daily and hourly variations in energy use are masked when consumption data are aggregated to the monthly level. With the availability of Advanced Metering Infrastructure (AMI) data, billing regressions can now be estimated at the hourly and even 15 minute intervals, rather than monthly level. With this advance comes the promise of potentially more accurate billing regression models. This paper provides a targeted comparison of billing regression model results estimated using monthly, daily and hourly consumption data. We use a fixed effects regression, which is a preferred model specification for many billing regression applications. The fixed effects model has the advantage of using indicator variables to control for both time and customer invariant factors, helping minimize bias and reducing the need for collecting additional data. Using the same model specification for each aggregation level, we assess the improvement in model fit and precision for key variables based on the shift from monthly to hourly data. The model is estimated using existing data from Southern California Electric's HVAC Quality Installation Program containing whole house metered data at the hourly level. This rich dataset presents an opportunity to test how billing models might be improved with the use of more granular consumption data. In addition to the model results, this paper will also provide recommendations for optimal model specification and data preparation. This paper will be of interest to evaluation practitioners that use billing regressions to estimate energy savings and are interested in enhancing models using AMI data.

Introduction

Historically, the evaluation community has largely relied on monthly energy bill data to estimate gross energy savings achieved by residential energy efficiency programs when using top-down regression analysis. An alternative to monthly bill data is data derived from Advanced Metering Infrastructure (AMI). AMI supports meters that measure and record usage data at a minimum, in hourly intervals, with some metering protocols collecting data at 15 minute intervals or finer. AMI is being rapidly rolled out across the United States with a recent report from the Department of Energy (DOE) estimating that approximately 30 percent of residences and 25 percent of commercial sites across the country, with over 50 percent of residences having AMI in territories of the Western Electricity Coordinating Council, as well as Texas and Florida (DOE 2014). As AMI becomes more prevalent, the evaluation community will have greater access to the rich data sources provided by AMI. One potential boon of AMI data is the additional variation in consumption data collected at the hourly or sub-hourly level that is masked in

traditional monthly billing data. This additional variability opens pathways to the potential for more accuracy in savings estimates attributed to energy efficiency programs.

This paper aims to provide a comparison of billing regression model results estimated using monthly, daily and hourly consumption data derived from a single AMI data source. Using the same model specification each aggregation level, we can assess the improvement in model fit and precision for key variables based on the shift from monthly to hourly data. We will use a fixed effects regression model, which is a common model specification for many billing regression applications. The model is estimated using AMI data collected from Southern California Edison customers participating in the HVAC Quality Installation (QI) Program. This program is designed to achieve energy and demand savings through the installation of replacement split or packaged HVAC systems in accordance with industry standards. These data contain whole house metered data at the hourly level.

Research Question

Hourly and daily energy consumption data can reveal significant variation in both energy use and weather data that is masked in monthly bill data. Figure 1, below presents an example month from the QI AMI dataset representing the variation of kWh consumption at different data aggregation levels. The green line representing kWh consumption averaged over the month – typical in monthly billing data – is a flat line representing the average daily consumption value for the month of July 2013. The red, daily kWh line exhibits more variation as it presents actual daily kWh consumption averages. Finally, the blue hourly consumption line illustrates the significant amount of variation in kWh consumption that occurs in an average household.

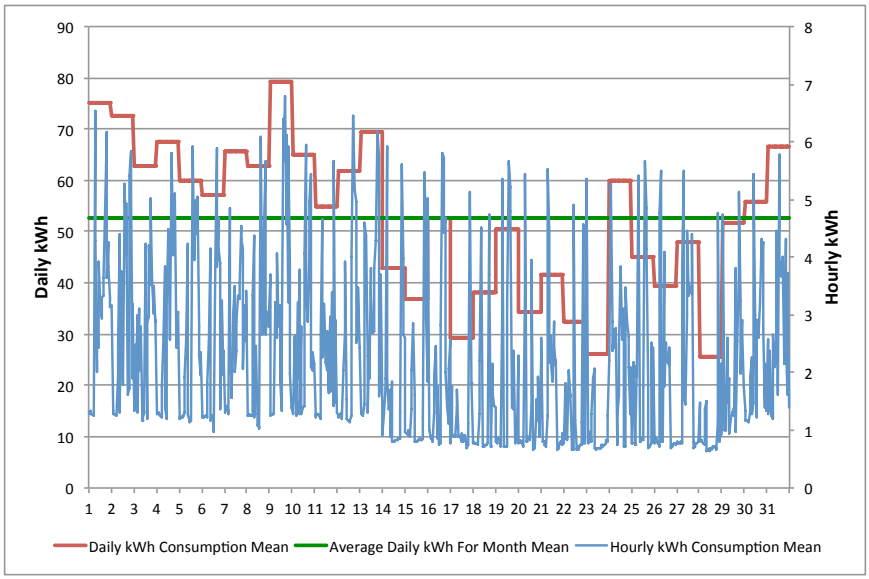

Figure 1 : Variation in Hourly, Daily and Monthly kWh Aggregations, July 2013

In this study, we seek to determine the relative benefit of analyzing AMI data at three aggregation levels: hourly, daily, and monthly, in order to account for this variation in data. Our key research questions are: do savings estimates change across levels of data?

- Are estimates at more disaggregated levels more precise?

Our hypothesis is that hourly and daily consumption data will provide more accurate savings estimates, and potentially reveal greater average savings attributable to the SCE QI program.

Methods

Data Sources

This analysis relies primarily on hourly AMI data and program participation data collected on residential customers that participated in SCE's Residential Quality Installation (QI) HVAC program between January 2012 and December 2014. The Residential QI program is a California statewide program that is designed to achieve energy and demand savings through the installation of replacement split or packaged HVAC systems in accordance with industry standards. The program aims to ensure that replacement HVAC systems are sized appropriately for a residence and are installed according to industry best practices (SCE 2013). Additionally, hourly weather station data sourced from the National Oceanic and Atmospheric Administration (NOAA) weather data were appended for each site.

Data Cleaning and Screens

In theory, AMI can provide high-quality, timely and reliable interval data, however, AMI is still susceptible to data collection errors and anomalies due to downtime in network communication, software errors or other factors. While utilities, including SCE, have meter data management systems with robust quality control mechanisms to validate data and estimate missing data, we conducted a detailed review of the QI customer AMI data to clean any potentially erroneous data from the modeling dataset. While the SCE QI customer AMI data proved to be of very high quality, the following cleaning steps were taken:

Sufficient Pre-Post Period Observations: While ideally a full year of data pre and post program intervention is available, this restriction would limit our modeling dataset to a small fraction of the available homes. We included only sites with at least nine month of data pre- and post- HVAC equipment installation that included at least one heating season month (July – September) in both the pre- and post- periods. [1] After this screen was imposed the analysis dataset includes 678 homes.

Missing Data: The data were search for missing hourly interval reads. Zero missing reads were identified.

Zero Reads: As noted in the Uniform Methods Project (UMP), zero reads are rare and sites with extensive electric zero reads should be removed. Zero reads accounted

[1] AMI data at the hourly or daily level introduce the possibility that a less data may be required for accurate savings estimation, however, for comparability across models in our analysis we maintained the same data screens for all aggregation levels.

for less than one-tenth of one percent of the total reads. We identified 42 days with 12 or more zero reads and these were removed from the analysis dataset.

Extreme Data: Days with high usage, in excess of 100kWh, were removed if the daily kWh was greater than 3 standard deviations from the specific home mean usage. This resulted in the removal of approximately 0.02 percent of days.

Table 1, below summarizes the data attrition attributable to the cleaning and screens we applied to the data. Analysis was conducted with more and less restrictive screens to test the sensitivity of the results; none altered the results or statistical significance of the results to a large degree.

Table 1. Data Attrition Through Data Screening

	All Data	Data Screened	Data Remaining	% of Total Screened
Hourly Observations	28,555,536	17,244,317	11,311,219	39.61%
Daily Observations	1,189,814	718,280	471,534	39.63%
Households	2,091	1,413	678	32.42%

Data Aggregation

Following data cleaning, we proceeded to create datasets at the three aggregation levels of interest, hourly, daily and monthly.

Hourly Dataset*:* The original hourly AMI data was manipulated to form a panel dataset suitable for analysis with each observation representing a single hour, day, home-record. Program data containing the HVAC equipment installation date was then appended and the pre- and post- installation periods defined for each household. Periods during which the installation occurred were flagged as blackout periods and not included in the analysis. At the hourly aggregation level, the day of installation was flagged as a blackout period. Hourly weather station data including actual average hourly temperature were retrieved from NOAA and appended to the hourly AMI data. We selected weather station data based on proximity to each observation home's zip code, matching climate zone, and availability of complete hourly data. The selection process resulted in hourly data for 95.5 percent of hourly observations; the remaining hourly weather data were interpolated by taking the mean of the preceding and following temperature reads. Accurate mean hourly temperature data allowed us to create heating and cooling degree variables at the hourly level. We computed hourly degree days by taking the difference between the average hourly temperature and a base temperature of 65^{O} F and dividing by 24, with hourly temperature less than 65^{O} F being heating hour and greater than 65^{O} F being cooling hours:

$$DD_{hourly} = \frac{1}{24} * (basetemp(65) - \overline{Temp}_{hour})$$

Daily Dataset*:* To aggregate to the daily level we simply take the daily sum of hourly kWh consumption, hourly HDD and hourly CDD, to get daily kWh consumption and daily HDD and CDD. The dataset is then limited to one row representing a single day, home record.

Monthly Dataset*:* Similarly, aggregation to the monthly level involves taking the sum of daily HDD and CDD for each month, and normalizing the resulting values to the average month length. For ease of comparability, rather than taking the sum of daily kWh, we calculate the average daily consumption (ADC) for each month. ADC is an equivalent variable to normalized monthly kWh and is the recommended consumption variable according to the UMP. The resulting dataset is limited to one observation representing a single month, home record.

Table 2 below presents a summary of the 3 datasets:

Table 2. Dataset Summary by Aggregation Level

	Observations	Households	Average kWh*	Average CDD*	Average HDD*
Hourly	11,311,219	678	1.10	.27	.23
Daily	471,534	678	26.54	7.65	6.53
Monthly	15,921	678	785.87	191.98	162.10

* Average kWh, CDD and HDD values are given at each aggregation level, hourly degree-days, daily degree-days and monthly degree-days.

It is important to note that aggregation of AMI data by this method allows the evaluator to create perfectly aligned monthly "bill" records that begin and end at the start and finish of each month. This eliminates many of the issues facing evaluators when dealing with traditional monthly billing records, such as off-cycle reads, billing adjustments, overlapping read intervals, and varying billing periods across households. In our analysis, we are comparing different levels of aggregation of AMI data only, and not comparing with traditional billing records.

Fixed Effects Model Specification

The fixed effects model is becoming a preferred model specification for many billing regression applications and is the model we chose to compare estimated savings from each data aggregation level. Pooled fixed effects regression combines all participants and time-periods in a single regression analysis and is an appropriate modeling approach when no comparison group available, as is the case with this analysis (NREL 2013). The benefit of the fixed-effects model is that it controls for unique time and customer invariant characteristics, or "fixed effects", within households, such as general levels of electricity use (i.e. a high usage or low usage household), home size and home occupancy, which could not otherwise be represented in the model. The fixed-effects model controls for these time and customer invariant characteristics through estimation of a household-specific constant term.

Models were estimated separately for each level of aggregation, hourly, daily and monthly. The fixed-effects model specification, detailed subsequently, is kept consistent across the modeling of each aggregation level. Due to the differing levels of aggregation, however, some of the specific variables are altered for each level of aggregation, specifically:

- The monthly fixed effects model takes the dependent variable, Average Daily kWh Consumption and independent variables total monthly heating degree-days and total monthly cooling degree-days.
- The daily fixed effects model takes the dependent variable, Actual Daily kWh Consumption and independent variables total daily heating degree days and total daily cooling degree days
- The hourly fixed effects model takes the dependent variable, Actual Hourly kW Consumption, and independent variables hourly heating degree-days and hourly cooling degree-days.
  The model is specified as follows:

$$kWh_{i,t} = \alpha_i + \beta_1(Post_{i,t}) + \beta_2(C_{i,t}) + \beta_3(H_{i,t}) + \beta_4(C_{i,t} * Post_{i,t}) + \beta_5(H_{i,t} * Post_{i,t}) + \sum_{j=6}^{16} \beta_j(M_t) + \varepsilon_{i,t}$$

Where:

$kWh_{i,t}$ = (Monthly Model) *Average* daily kWh consumption in month t for customer i.
(Daily Model) *Actual* daily kWh consumption in day t for customer i.
(Hourly Model) *Actual* hourly kWh consumption in hour t for customer i.

$Post_{i,t}$ = (Monthly Model) A dummy variable indicating post HVAC installation month t for customer i.
(Daily Model) A dummy variable indicating post HVAC installation day t for customer i.
(Hourly Model) A dummy variable indicating hour in post HVAC installation day t for customer i.

$C_{i,t}$ = (Monthly Model) Total monthly cooling degree days based on a base temperature of 65°F in month t for customer i.
(Daily Model) Daily cooling degree days based on a base temperature of 65°F.
(Hourly Model) Hourly cooling degree days based on a base temperature of 65°F.

$H_{i,t}$ = (Monthly Model) Total monthly heating degree days based on a base temperature of 65°F in month t for customer i.
(Daily Model) Daily heating degree days based on a base temperature of 65°F.
(Hourly Model) Hourly heating degree days based on a base temperature of 65°F.

$C_{i,t} * Post_{i,t}$ = Interaction between cooling degree day variable and post period indicator.

$H_{i,t} * Post_{i,t}$ = Interaction between heating degree day variable and post period indicator.

M_t = Set of dummy variables for each month excluding January.

$\beta_1,...,\beta_j...$ = Coefficients to be estimated in the regression model.

α_i = Household specific constant.

$\varepsilon_{i,t}$ = Random error term, assumed to be normally distributed.

The variables of interest that capture the savings estimation for each model aggregation level are all terms including the *Post* variable, *Post, C*Post* and *H*Post*. The coefficient on the *Post* variable can be interpreted as the average change in consumption attributable to a household in the post-installation period. The coefficient on the *C*Post* variable can

be interpreted as the average change in consumption attributable to a household in the post-installation period due to an increase of one cooling degree-day in that period. Likewise, the coefficient on the *H*Post* variable can be interpreted as the average change in consumption attributable to a household in the post-installation period due to an increase of one heating degree-day in that period. To calculate the average energy savings based on the regression results, the following equation is used:

$$Avg\Delta kWh_{i,t} = \beta_1 + \beta_4(\overline{C}) + \beta_5(\overline{H})$$

Estimates of the standard error of this transformation are calculated using the Delta Method.

The following section presents the results of the analysis.

Results

Table 3 presents the estimation results for Model 1, the monthly regression model. The coefficients of interest with respect to energy savings attributable to the HVAC QI program are β_1, β_4, and β_5. Each of these coefficients is statistically significant at the 1 percent level. While the coefficient β_1 is not of the expected sign, the coefficients β_4, and β_5 are and the transformation to estimated savings using the equation detailed in the previous section results in statistically significant savings of 2.42 kWh per day or 8.91 percent.

Table 3. Model 1 - Monthly Fixed Effects Regression

Model Summary	
Daily kWh Mean	27.17
ADC Standard Deviation	18.12
Number of Households	678
Number of Observations	15,921
Adjusted R-Squared	.497
Estimated Savings (95% CI)	2.42 $\pm$ 0.246 kWh (8.91% $\pm$ 0.91%)

Variable	Coefficient (β)	Standard Error	t-statistic	Sig. (p-value)
(β_1)Post (Month)	**2.433**	**0.337**	**7.224**	**<0.001**
(β_2)C	0.043	0.001	46.719	<0.001
(β_3)H	0.007	0.001	7.791	<0.001
(β_4)Post*C	**-0.019**	**0.001**	**-20.968**	**<0.001**
(β_5)Post*H	**-0.007**	**0.001**	**-6.734**	**<0.001**
(β_6)Feb	-0.672	0.289	-2.320	0.020
(β_7)Mar	-1.422	0.314	-4.533	<0.001
(β_8)Apr	-2.168	0.340	-6.382	<0.001
(β_9)May	-1.316	0.386	-3.408	0.001
(β_{10})Jun	2.370	0.432	5.488	<0.001
(β_{11})Jul	6.924	0.498	13.903	<0.001

(β_{12})Aug	5.691	0.486	11.721	<0.001
(β_{13})Sep	4.917	0.464	10.602	<0.001
(β_{14})Oct	-1.900	0.355	-5.357	<0.001
(β_{15})Nov	-0.551	0.296	-1.862	0.063
(β_{16})Dec	3.020	0.290	10.398	<0.001

Table 4 presents the estimation results for Model 2, the daily regression model. Again, the coefficients of interest with respect to energy savings attributable to the HVAC QI program are β_1, $\beta_4, and\ \beta_5$. Similarly, the coefficients are statistically significant at the 1 percent level. While the coefficient β_1 is not of the expected sign, the coefficients $\beta_4, and\ \beta_5$ are and the transformation to estimated savings using the equation detailed in the previous section results in higher savings than the monthly model of 2.63 kWh per day or 9.68 percent. Additionally, the precision of the savings estimate is higher with lower standard errors on the coefficients of interest and a tighter confidence interval. However, the 95 percent confidence interval for the daily estimated savings falls within the 95 percent confidence interval of the monthly regression results, indicating that while the savings difference is not statistically significant, the accuracy of the savings results is higher.

Table 4. Model 2 - Daily Fixed Effects Regression

Model Summary	
Daily kWh Mean	27.17
ADC Standard Deviation	18.12
Number of Households	678
Number of Observations	471,534
Adjusted R-Squared	.397
Estimated Savings (95% CI)	2.63 ± 0.062 kWh ($9.68\% \pm 0.23\%$)

Variable	**Coefficient (β)**	**Standard Error**	**t-statistic**	**Sig. (p-value)**
(β_1)Post (Day)	1.564	0.071	22.087	<0.001
(β_2)C	1.374	0.005	295.976	<0.001
(β_3)H	0.274	0.004	61.270	<0.001
(β_4)Post*C	-0.516	0.006	-90.968	<0.001
(β_5)Post*H	-0.123	0.006	-19.782	<0.001
(β_6)Feb	-0.937	0.075	-12.520	<0.001
(β_7)Mar	-1.696	0.079	-21.341	<0.001
(β_8)Apr	-2.132	0.081	-26.442	<0.001
(β_9)May	-1.006	0.081	-12.358	<0.001
(β_{10})Jun	2.185	0.085	25.624	<0.001
(β_{11})Jul	6.858	0.091	75.214	<0.001
(β_{12})Aug	5.664	0.090	63.275	<0.001
(β_{13})Sep	4.513	0.088	51.030	<0.001
(β_{14})Oct	-1.562	0.078	-20.044	<0.001

(β_{15})Nov	-0.475	0.074	-6.415	<0.001
(β_{16})Dec	2.902	0.073	39.779	<0.001

Finally, Table 5 presents the estimation results for Model 3, the hourly regression model. Similarly, the coefficients are statistically significant at the 1 percent level. In this case, the coefficient β_1 is of the expected sign, as are the coefficients $\beta_4, and\ \beta_5$. The estimated savings derived from the equation detailed earlier are again slightly higher than the daily model at 2.64 kWh per day or 9.74 percent. Again, the precision of the savings estimate is higher with lower standard errors on the coefficients of interest and a tighter confidence interval. In addition, the 95 percent confidence interval for the hourly estimated savings falls within the 95 percent confidence interval of the monthly and daily regression results.

Table 5. Model 2 - Hourly Fixed Effects Regression

Model Summary	
Daily kWh Mean	27.17
ADC Standard Deviation	18.12
Number of Households	678
Number of Observations	11,311,219
Adjusted R-Squared	.397
Estimated Savings (95% CI)	2.64 $\pm$ 0.013 kWh (9.74% $\pm$ 0.09%)

Variable	**Coefficient (β)**	**Standard Error**	**t-statistic**	**Sig. (p-value)**
(β_1)Post (Hour)	-0.023	0.001	-26.931	<0.001
(β_2)C	1.148	0.001	1007.624	<0.001
(β_3)H	0.100	0.001	89.695	<0.001
(β_4)Post*C	-0.290	0.002	-189.973	<0.001
(β_5)Post*H	-0.013	0.002	-7.839	<0.001
(β_6)Feb	-0.042	0.001	-32.388	<0.001
(β_7)Mar	-0.084	0.001	-61.516	<0.001
(β_8)Apr	-0.105	0.001	-76.320	<0.001
(β_9)May	-0.056	0.001	-41.773	<0.001
(β_{10})Jun	0.087	0.001	65.025	<0.001
(β_{11})Jul	0.291	0.001	211.931	<0.001
(β_{12})Aug	0.241	0.001	176.993	<0.001
(β_{13})Sep	0.188	0.001	137.867	<0.001
(β_{14})Oct	-0.082	0.001	-63.334	<0.001
(β_{15})Nov	-0.028	0.001	-22.061	<0.001
(β_{16})Dec	0.132	0.001	104.647	<0.001

Figure 2 below presents a graphical representation of the savings estimates for each model. Energy savings calculated from aggregated monthly data differed by

approximately 0.75% from estimates made using hourly and daily data. While this difference is small and not statistically significant, the upward trend in savings suggests that using finer levels of data can capture additional savings not captured by monthly data. The results do show that the accuracy of savings results derived from more granular, daily or hourly AMI data is likely to be higher.

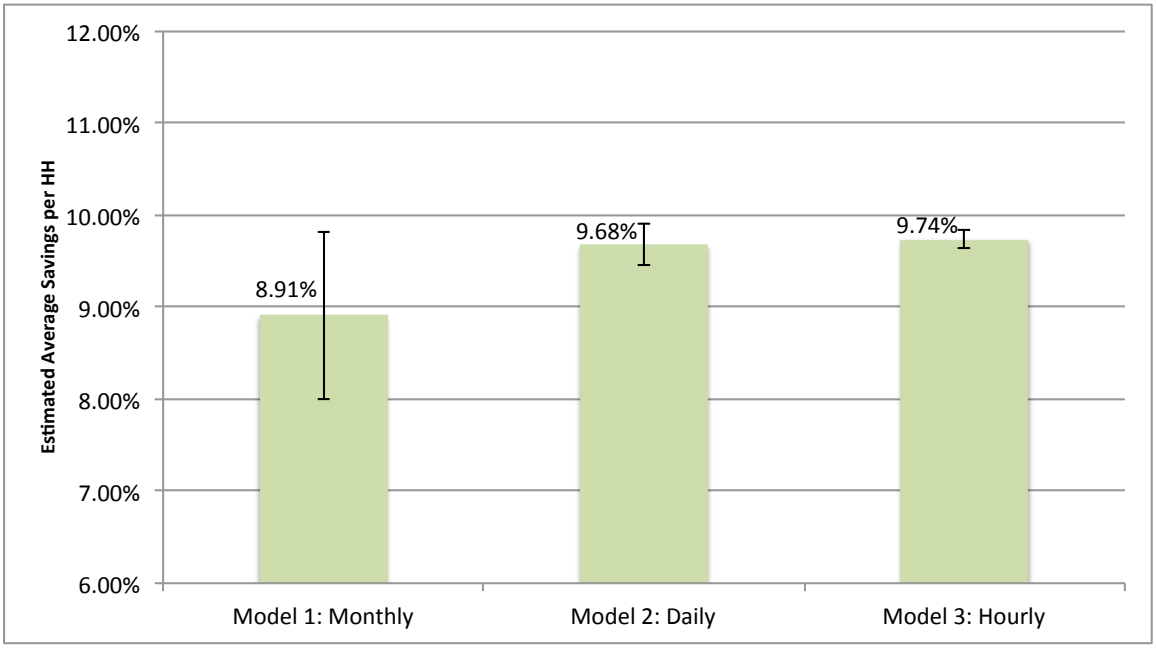

Figure 2. Comparison of Savings Estimates

Discussion

The results of this analysis suggest there is potential for improving savings estimates by using AMI data; AMI data can deliver more precise savings estimates and potentially identify additional savings not uncovered by monthly data analysis – more may be better. However, this analysis should be of comfort to the evaluation community, as it also suggests that estimating energy savings based on monthly data is suitable and provides reasonable estimates of gross energy savings.

A significant challenge in the estimation of savings using hourly or finer AMI data is the computational resources, time and cost required to handle very large datasets. The results of this analysis may indicate that using accurate AMI data aggregated to the monthly (or daily) level may be sufficient for estimating gross energy savings, meaning utilities and evaluators could save valuable resources that would be required to estimate savings using hourly or finer data.

While AMI data may not greatly change estimates of *overall* energy savings, it does open up important new pathways in energy efficiency research. Consumption data at the hourly or finer level will allow evaluators to assess the quality of energy savings. Since utilities must maintain sufficient generation infrastructure to deliver energy to meet

peak demand, savings at peak-load times are more valuable than savings at times of surplus energy. Identifying programs that achieve savings during peak times could reduce the need for new energy infrastructure. Evaluators can potentially use AMI data to estimate savings at different times of day to estimate savings impacts in peak demand periods as well as overall. This analysis is being conducted in conjunction with additional analysis that is investigating the potential of AMI data in estimating savings at specific times of day.

Limitations of Analysis

As noted previously, this analysis compares different levels of aggregated AMI data only. While this process provides valuable insight into the effect of disaggregation of data on estimation of savings, it does not directly compare the use of AMI data to traditional billing data. Traditional consumption data derived from bills is inherently error prone, with greater exposure to human error, as well as temporal inconsistencies across billing periods and households. Without access to traditional billing records for the same households, we are unable to truly compare AMI data with traditional billing data. It may be reasonable however to assume that comparing AMI data with actual monthly bill data will result in greater differences in estimated savings than those shown in this analysis.

Areas for Future Research

Time of Day Savings - while AMI data may not change the way we calculate overall energy savings, it creates the opportunity to evaluate savings at different times of day. Hourly and daily consumption data could allow evaluators to assess the quality of energy savings, arguably more important than aggregate savings. Future research in the estimation of regression models, including fixed effects models, to include estimates of hourly savings would be of great value to the energy efficiency community.

AMI vs. Monthly Billing Data - future research could compare aggregated monthly AMI data to traditional monthly billing data for the same time period. This analysis could establish whether using aggregated monthly AMI data provides a significant benefit over billing data in calculating overall energy savings.

Conclusion

This paper presents the results of a unique analysis of savings estimates developed from different aggregation levels of hourly interval AMI data collected from customers that participated in SCE's HVAC QI program from December 2012 to December 2014. The analysis estimates energy savings attributable to the program using monthly, daily and hourly data aggregation levels to determine if, with greater granularity of consumption data comes the promise of potentially more accurate billing regression models. The results of the analysis indicate an increase in savings estimates as data is disaggregated from monthly to daily and again to hourly levels with estimated savings at the monthly level of 8.91 percent, at the daily level of 9.68 percent and the hourly level of 9.74 percent. Additionally with each level of disaggregation, the standard errors of key

variables of interest, as well as the overall savings estimates reduced indicating that with finer levels of data comes more accurate savings estimations.

References

Dubin, J. A., and V. Gamponia, 2007: Mid-Range, Average, and Hourly Estimates of Heating Degree Days: Implications for Weather Normalization of Energy Demand. The Energy Journal, April 2007.

Federal Energy Regulatory Commission. 2014. *Assessment of Demand Response and Advanced Metering.* Staff Report.

Southern California Edison. 2013. *Customer Energy Efficiency and Solar Division Program Implementation Plans 2013 – 2014*.

Agnew, K. & Goldberg, M. (2013). *Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol*. Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. Golden, CO: National Renewable Energy Laboratory, NREL/SR-7A30-53827 April 2013.