Challenges of Estimating Hourly Baselines for Residential Customers

Collin Elliot, Jean Shelton, and Dave Hanna, Itron, San Diego, CA

ABSTRACT

The expansion of smart grid infrastructure has offered new opportunities for utilities to implement demand response programs for their residential customers. The standard approaches for estimating the baseline profiles (BLPs) used to determine impacts, however, come primarily from non-residential programs, and their applicability to residential customers presents two major challenges. First, the number of participants in residential programs – potentially millions – can render standard methods computationally impractical. Second, the greater variability of residential loads can make the standard approaches insufficiently accurate for producing reasonable estimates of program impacts. This paper summarizes analysis of interval data for residential participants in SCE's SmartConnect program to assess, first, whether the standard BLP approaches are sufficiently accurate for estimating impacts for individual residential customers and, second, whether the use of aggregate customer load models offer a reliable and more efficient alternative.

Introduction

The expansion of smart grid infrastructure has offered new opportunities for utilities to implement demand response programs for their residential customers. Following the installation of residential smart meters, utilities are able to track customers' energy use and offer incentives to customers who are enrolled in a demand response program and curtail their energy use during peak demand periods. Typically utilities call event days -- or days when energy demand has the potential to surpass energy supply due to abnormal weather patterns, such as heat waves – to signal when program participants can receive an incentive for reducing their energy use. To determine each customer's true savings on an event day and provide incentives that accurately value the customer's savings, utilities must compare each program participant's event day energy use to their individual baseline use on a non-incentivized peak day. The standard approaches for estimating the baseline profiles (BLPs) used to determine impacts, however, come primarily from non-residential programs, and their applicability to residential customers presents two major challenges. First, the number of participants in residential programs – potentially millions – can render standard non-residential methods computationally impractical. Second, the greater variability of residential loads can make the standard approaches insufficiently accurate for producing reasonable estimates of program impacts.

This paper assesses the performance of different BLPs for residential customers, using a sample of participants in Southern California Edison's (SCE's) Save the Power Day Incentive (SPDI) program. As the AMI smart meters are installed and cut over to operation, residential (and small commercial) customers are automatically enrolled or opted-into the SPDI program or peak time rebate rate. The SPDI program will have well over three million participants in the near future. The study presented in this paper focused on evaluating the accuracy of baselines generated for each individual household in the sample using non-residential BLPs. Individual baselines were generated using both "representative day" approaches, where the BLP is based on an average of non-event days prior to the actual event, and regression analyses, where the hourly usage is estimated as a function of weather and calendar variables ("dummy" variables for day of week, etc.). In addition to account-level regression analysis, this study also included a regression-based BLP using aggregated hourly data from the same sample of accounts used in the account-level analysis. This secondary regression model was evaluated to address whether it is better to estimate baselines for individual accounts and then aggregate the results, or to aggregate hourly data and then estimate average impacts. One clear disadvantage with individual models is the amount of data processing required. Nevertheless, there are some advantages to estimating account-level baselines. For one, individual baselines often benefit from more geographically-specific weather data and they should more accurately capture any site-specific calendar type effects (e. g. a day of the week where the occupant is consistently away).

Data Sources

The analysis presented in this paper relied on four data sources:

- ***Interval Data***: Hourly kWh readings for more than 80,000 accounts for varying data ranges in 2011. The interval data are for residential customers in Foothill, Whittier, and Wildomar meter districts within SCE's territory.
- ***Customer Information System (CIS) Data***: Information on rate codes, climates zones, etc. used for developing different customer strata.
- ***Weather Data***: Hourly temperature, humidity, and other weather variables for six climate zones for January through October of 2011.
- ***CA ISO Data***: Hourly system load information for 2011 for SCE service territory and statewide.

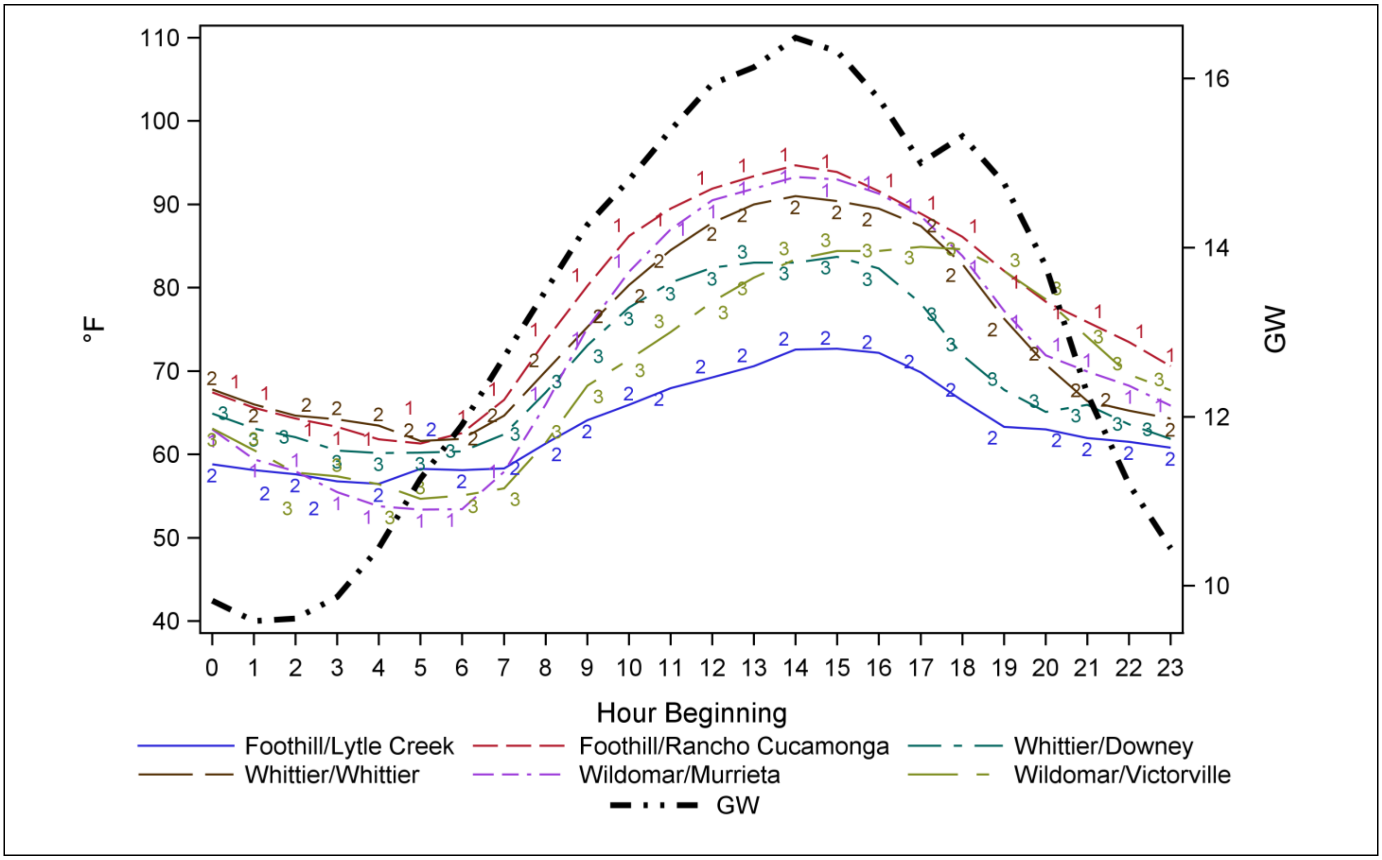
Sample Frame and Sample Selection

Given the resources necessary to estimate the various baselines for the full population of available SPDI accounts, the CIS and usage data were used to develop a sample frame from which to select a random sample. The strata for this study were geography (meter district), enrollment in the California Alternate Rates for Energy (CARE) program, and usage category (based on terciles for annual consumption) for a total of 18 separate strata. The purpose for these strata is to differentiate between customers in different climate zones or usage patterns in a way that will identify whether BLPs work better for some groups of customers than for others. A total of 100 accounts per strata were randomly selected for a total final sample of 1,800 residential accounts.

Proxy Event Day Analysis

The objective of the proxy event day analysis was to select a day in each month that would be similar to the type of day when an event might be called. Events are typically called on hot days, where AC load results in strains on the distribution system. An initial approach used weather data to identify the hottest days – using several definitions – but a minor issue was inconsistency across the weather stations in terms of which days were hottest. Rather than select the days based on weather, the proxy event days were based on peak days in each month for the CAISO hourly load data for SCE's service territory. This definition of proxy event days allowed all strata to rely on a common criterion for event days and the SCE peak loads generally correspond to those days with high temperatures. As an example, Figure 1 shows the SCE system load for one of the proxy event days along with the hourly temperatures for six combinations of district and weather station. The district and weather station series are annotated with the monthly rank in terms of average daily temperature. As the figure shows, the day represented at least the third hottest day in May and for two of the weather stations it was the hottest day in the month.

Figure 1: May 4th 2011 SCE System Peak with Weather by District/Weather Station



Representative Day Baseline Calculations

Having selected a set of proxy event days, the next step in the study was the estimation of the individual BLPs. The first set of BLPs was the following five different types of representative day baselines:

10-Day Baseline: Average of the ten previous non-holiday non-event weekdays (NHNEWD) before an event.
Top 3 of 10 Baseline: Average of the top three of the previous ten NHNEWDs prior to the proxy event day, where the days are sorted based on their consumption during the event timeframe.
Top 3 of 5 Baseline: Average of the top three of the previous five NHNEWDs prior to the proxy event day, where the days are sorted based on their consumption during the event timeframe.
Top 3 of 10 Baseline, Adjusted: The Top 3 of 10 Baseline adjusted based on the pre-event load.
Top 3 of 5 Baseline, Adjusted: The Top 3 of 5 Baseline adjusted based on the pre-event load.

The adjustments for the final two baselines are multiplicative factors based on the two hours prior to an event period, which for this analysis were assumed to begin at 2:00 p.m. Each multiplicative factor is defined as the ratio of the actual to the predicted (BLP) load in the two hours prior to the event period. The average of the two hours prior to the event on the event day will be divided by the average of the same two hours for the BLP. The adjusted BLP is simply the BLP in each hour multiplied by the adjustment factor.

Two positive characteristics of the representative day BLPs are that they are relatively simple to calculate and easy to understand. The idea that a customer's load on an event day will be the average of some set of similar non-event days is generally intuitive. Yet in spite of this, there are many different ideas about which set of non-event days will best replicate the load profile. As described above, this assessment looked at three different ways of selecting the NHNEWDs.

Figure 2 and Figure 3 present examples of the three types of representative-day BLPs along with the loads for the ten NHNENWDs that contributed to their calculation. Each series for the individual days is labeled so that it is clear to which of BLPs it contributed. Additionally, the event hours are outlined by a dashed box, which is the period of interest. Figure 2 is for a customer that exhibited relative uniform load in the contributing days. In contrast, Figure 3 is based on a customer with substantially higher variability. What these examples illustrate is that for customers with highly variable load, the days selected for the BLP can result in dramatically different BLPs. In Figure 2, there is one day that makes the 3 of 10 baseline slightly higher than the 3 of 5 day BLPs, but only slightly so. In contrast, there are stark differences in three BLPs in Figure 3. The 3 of 10 BLP is affected by one particularly high day that occurred outside of the five-day range. Conversely, the 10 day baseline is substantially lower than the other two due to three days of particularly low consumption. Which of these baselines will perform best for the variable load customer, then, depends on the load on the proxy event day. In addition, the "Other 10 day" lines in Figure 2 and Figure 3 clearly illustrate the variability of residential loads, even for the relatively constant load represented in Figure 2.

Figure 2: Example of Baseline and Contributing Days, Low Load Variability

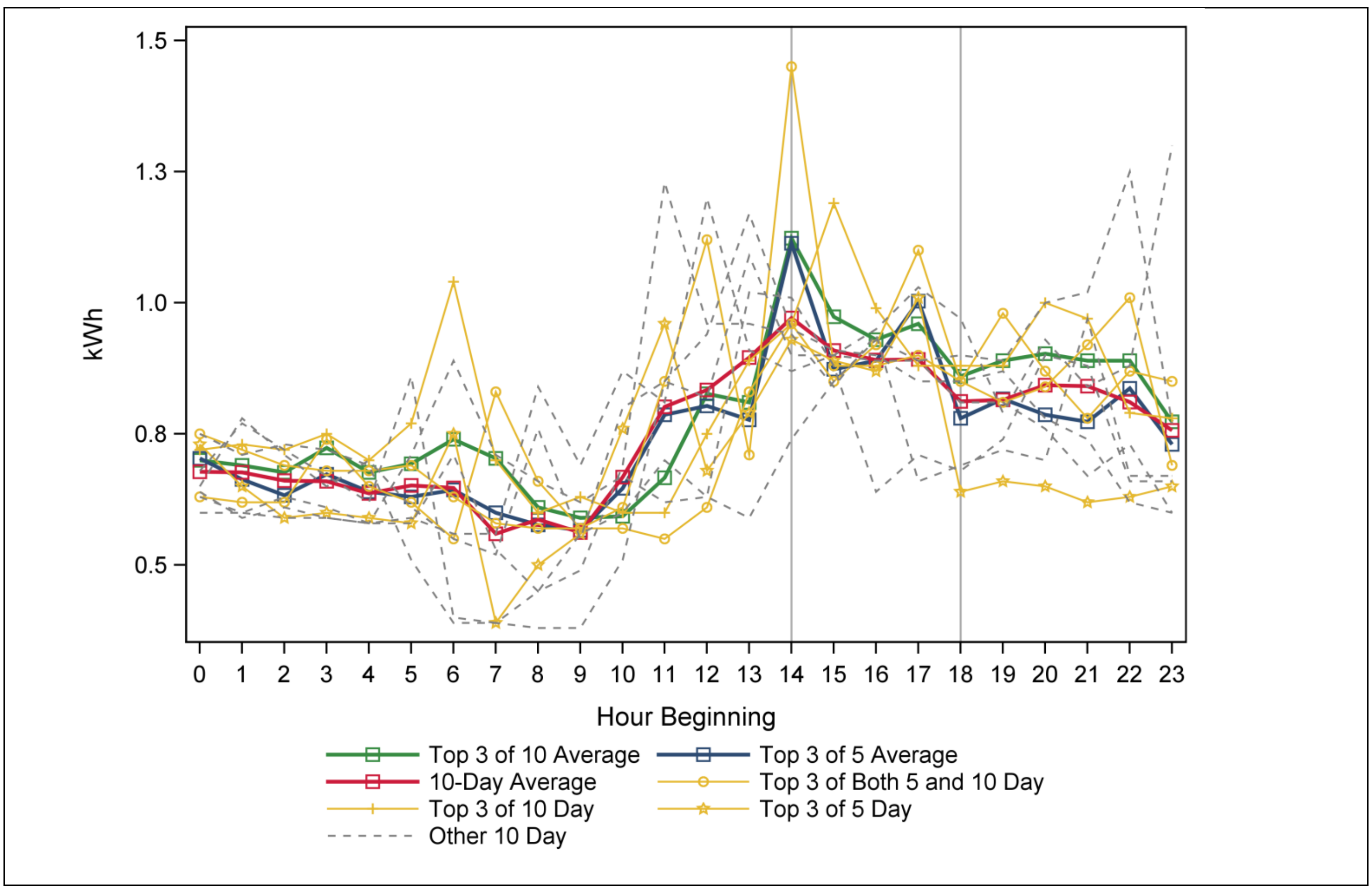
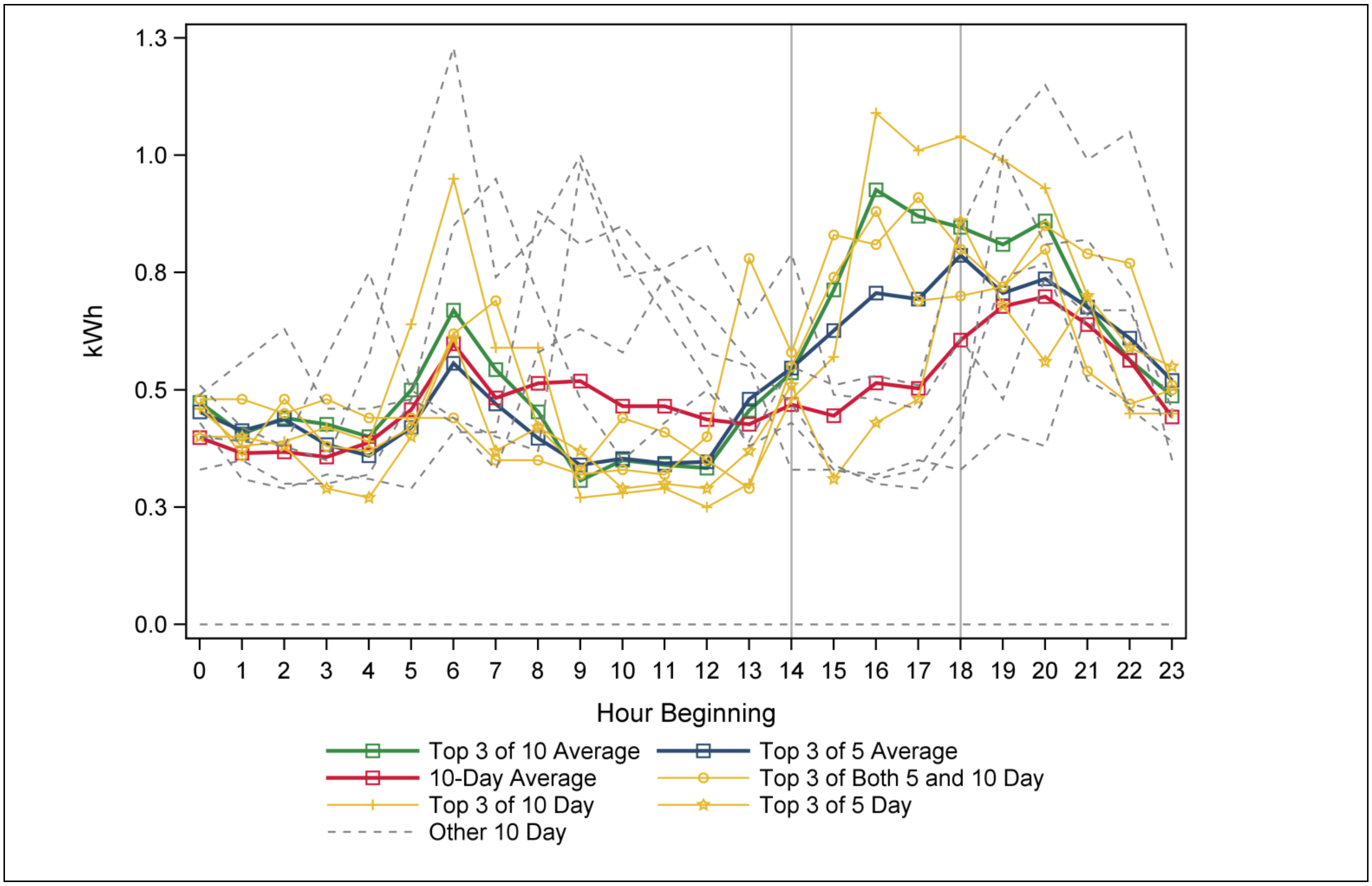


Figure 3: Example of Baseline and Contributing Days, High Load Variability



Regression Baseline Methodology

The next set of BLPs included in this assessment was based on account-level regression models. The proxy event days were excluded from the model and the hourly load was estimated using a variety of weather and calendar variables. The first BLP in the set was based on the predicted load for the proxy event days. The second BLP applied the pre-event adjustment in a manner identical to that used for the representative day baselines.

For the regression baseline, weather data were combined with hourly interval data and the model shown in Equation 1 was estimated separately for every account using all data for the non-proxy event days. With the exception of some cases that had slightly different ranges of dates that excluded some of the calendar-related variables, the same model was applied to all accounts.

Equation 1

$$kWh_h = \propto + \sum_{m=1}^{9}(\beta_m \times m) + \sum_{h=1}^{23}(\beta_h \times h) + \sum_{d=1}^{7}\sum_{h=1}^{23}(\beta_{hd} \times h \times d) + \beta_{CDH} \times CDH + \beta_{HDH} \times HDH + \beta_{weekendHoliday} \times weekendOrHolidy + \beta_{preEventLoad} \times preEventLoad + \beta_{lagCDH} \times lagCDH + \varepsilon$$

Where…

kWh_h = kWh for hour h

α_i = account intercept

β_m = parameters for series of monthly dummy variables for January through September

β_h = parameters for series of hourly dummy variables
β_{hd} = parameters for series of interaction of day of week and hourly dummy variables
β_{CDH} = parameter for cooling degree hours (base 65)
β_{HDH} = parameter estimate for heating degree hours (base 65)
$\beta_{weekendHoliday}$ = parameter estimate weekend or holiday dummy variable
$\beta_{preEventLoad}$ = parameter estimate for load two hours prior to event time frame
β_{lagCDH} = parameter for a 24 hour lag of the CDH
ε = error for account

Regression models are not always effective at predicting energy consumption. Additionally, the estimation of BLPs for 1,800 accounts does not allow for easily customized model specifications. As a result, there will be among the many accounts, models that simply will not fit the data well. Table 4 shows the minimum, maximum, and average R^2 statistics by strata. The R^2 statistic can range from 0 to 1, where one indicates that the independent variables in the model explain 100% of the variation in the dependent variable. For the models estimated for this analysis, the mean R2 values are relatively high, which is indicative of models that should estimate the hourly load fairly well. Some customers have loads that lend themselves to regression modeling, as shown by some of the very high maximum R2 values, but at the same time there are also some sites with load profiles that are not explained by the independent variables in the regression model. There is a clear relationship between the usage tercile and the model fit, which is likely due to the presence of larger air conditioning systems that exhibit the expected weather sensitivity.

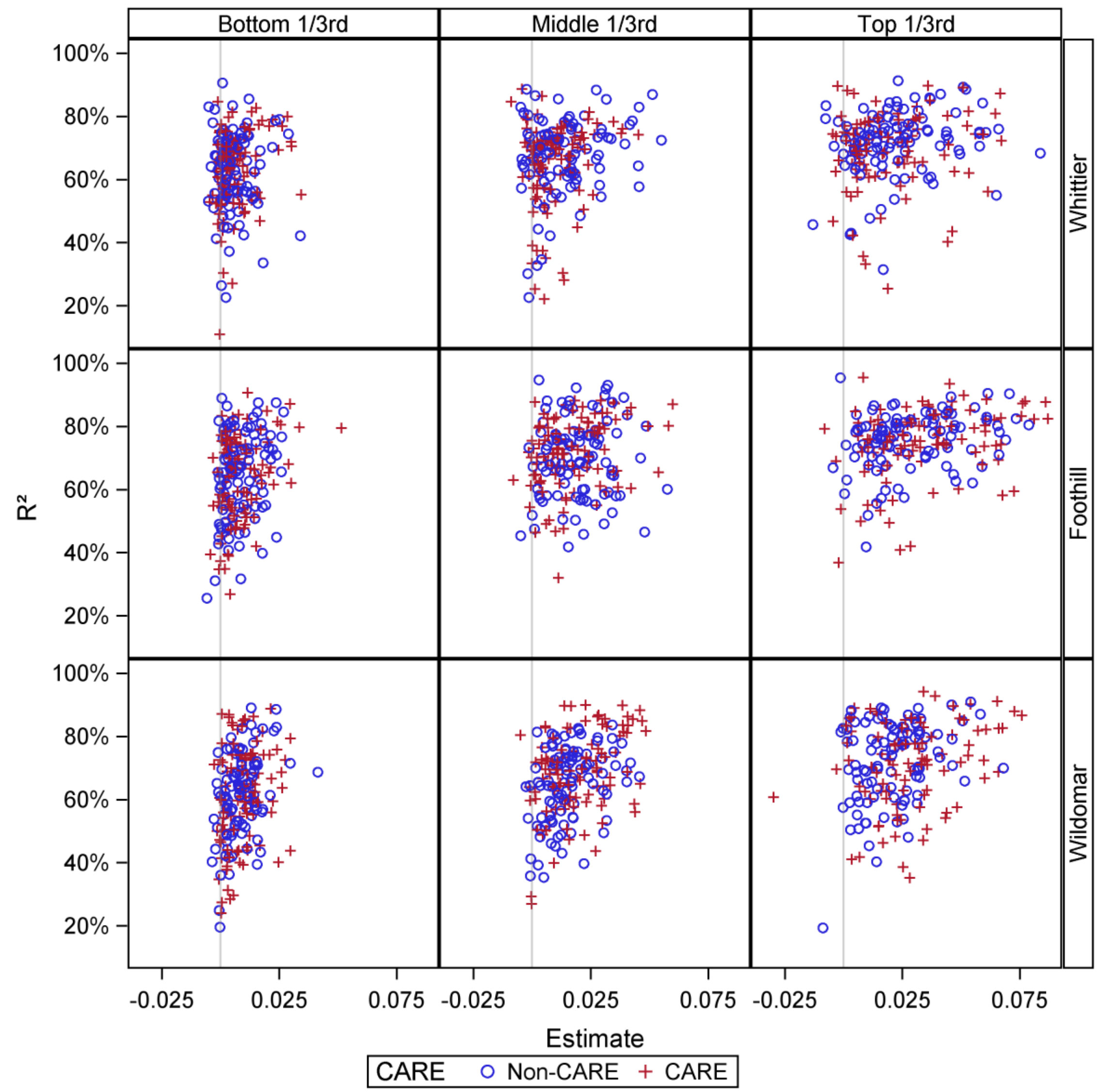
Table 1: Model Fit Summary for Regression Baselines

District	CARE Designation	Usage Tercile	Minimum R^2	Mean R^2	Maximum R^2
Whittier	Non-CARE	Bottom 1/3rd	0. 226	0. 625	0. 906
		Middle 1/3rd	0. 227	0. 674	0. 886
		Top 1/3rd	0. 314	0. 716	0. 913
	CARE	Bottom 1/3rd	0. 110	0. 629	0. 846
		Middle 1/3rd	0. 221	0. 651	0. 886
		Top 1/3rd	0. 254	0. 695	0. 897
Foothill	Non-CARE	Bottom 1/3rd	0. 256	0. 645	0. 890
		Middle 1/3rd	0. 418	0. 700	0. 947
		Top 1/3rd	0. 419	0. 757	0. 954
	CARE	Bottom 1/3rd	0. 269	0. 635	0. 907
		Middle 1/3rd	0. 320	0. 715	0. 882
		Top 1/3rd	0. 368	0. 747	0. 955
Wildomar	Non-CARE	Bottom 1/3rd	0. 196	0. 613	0. 890
		Middle 1/3rd	0. 353	0. 649	0. 836
		Top 1/3rd	0. 193	0. 719	0. 909
	CARE	Bottom 1/3rd	0. 240	0. 614	0. 888
		Middle 1/3rd	0. 269	0. 687	0. 900
		Top 1/3rd	0. 353	0. 717	0. 942

The model specification included dozens of variables, but the parameters for the weather variables are of primary interest because they should have a consistent relationship with the hourly consumption and they give an idea of how many of the accounts show sensitivity to weather. As a way of exploring the results for the weather variables and examining their relationship with how well the models fit the data, Figure 4 shows a scatter plot of the model R^2 and the parameter estimates for cooling degree hours. The plot is divided by each stratum and district and the points in the plot are represented by different colors for CARE and non-CARE customers. This configuration makes all of the effects for the strata dimensions discernible. Additionally, the plots show a vertical line at zero, which helps to show how many of the weather parameters were negative, which is generally a sign of models that do not work well.

The first observation in Figure 4 is the clear influence of the usage group on the relationship between cooling load and model fit. Logically, the parameter estimates for CDH grow larger as moving from the bottom to the top of the usage strata, but there is also a pattern of improved model fit. For the bottom usage stratum, the points form more of a vertically oriented cluster with a fairly even distribution along the axis showing the R^2. For the middle and top usage strata, the clusters of points shift to more horizontal alignments at higher points on the vertical axis. This pattern is more evident for the Foothill district, which is associated with higher temperatures. There is no clear sign that CARE designation has an influence.

Figure 4: Scatter Plots of Model R^2 by Parameter Estimate for CDH



While the review of the overall model fit and parameters is important, ultimately it says nothing about how well the baselines actually worked. To assess the actual BLPs relative to the actual load, we looked at one metric to quantify the accuracy and another to measure the bias in the BLPs. For accuracy, the metric is the absolute percent error (APE), which is the absolute percent difference between the estimated BLP and the actual load on the proxy event days. For bias, the metric is simply the percent error, which will show whether the differences between BLP and actual tend to be more positive or negative.

Table 2 presents a summary of the accuracy and bias metrics across all strata and proxy event days for the seven different account-level BLPs. Based on the median of the APE, which is used to avoid the influence of large outliers, none of the models are accurate. The adjusted regression BLP has the "best" accuracy, with a median APE of 23.2%, but this is only minimally better than the adjusted three-of-five day BLP. A review of the 95th percentiles shows that all of the approaches result in BLPs that can fit quite poorly, though it is interesting that the straight ten-day BLP results in a smaller number of cases with really poor fit, as suggested by its markedly lower 95th percentile APE.

Table 2: Baseline Metric Summary Across all Strata and Proxy Events

Baseline	Accuracy - APE			Bias - PE			
	5th Percentile	Median	95th Percentile	5th Percentile	Mean	Median	95th Percentile
10-Day	2. 7%	29. 4%	89. 5%	-87. 3%	4. 9%	14. 2%	72. 2%
Top 3 of 10	2. 1%	29. 4%	200. 5%	-200. 5%	-38. 5%	-10. 6%	56. 9%
Top 3 of 10, Adjusted	2. 1%	25. 4%	163. 1%	-163. 1%	-30. 2%	-10. 0%	55. 1%
Top 3 of 5	2. 1%	26. 4%	133. 3%	-133. 3%	-14. 8%	1. 9%	65. 5%
Top 3 of 5, Adjusted	1. 8%	23. 3%	111. 9%	-111. 9%	-11. 4%	0. 3%	63. 4%
Regression	2. 5%	26. 2%	148. 2%	-147. 9%	-20. 1%	-0. 7%	51. 2%
Regression, Adjusted	1. 8%	23. 2%	115. 2%	-114. 9%	-11. 9%	0. 3%	54. 5%

With respect to bias, PE values greater than zero indicate that the BLPs are lower than the actual load, and therefore underestimated. Conversely, values below zero indicate that the BLP overestimates load. If the median or average PE is substantially greater than or less than zero, it indicates that the models have a tendency to either over- or underestimate load. Based on the median PEs, some of the BLPs perform reasonably well. For example, both regression BLPs and the adjusted three-of-five day BLP have median PE values close to zero, meaning that approximately half of the models are overestimating and half are underestimating load. While the median values suggest little bias, it is worth noting that the average PE is substantially lower than zero, which means that there are some BLPs that substantially overestimating load.

Aggregated Regression Baselines

Given the limited reliability of account level baselines, the study next looked at whether using a regression model based on aggregated loads could provide more reliable means of estimating average customer loads. For this assessment, an hourly regression model was applied to the aggregated loads for accounts in the 18 strata used for sample design, with weather variables based on averages weighted by the customer load. As with the account-level baselines, a pre-event adjustment was applied to the aggregate regression BLPs to form a second set of adjusted aggregate BLPs.

While the aggregate BLPs could be assessed on their own, to compare them with the account-level baselines, the account-level regression BLPs were summarized by strata and proxy event to allow for a comparison with the aggregated baselines. The question behind this analysis is whether modeling on aggregated load results in a better baseline than one would get by simply averaging the account-level baselines.

Table 3 provides a summary by proxy event and overall for the APE across the 18 strata for the four aggregate BLPs, with the most accurate APE in green and the least accurate in red. While there is considerable variability by the proxy event day, the overall best BLP in terms of accuracy was the adjusted aggregate regression. This baseline's overall median absolute percent error (APE) of 5.9% is more than three percentage points lower than the account regression, which was the second best BLP. It is interesting to note that without any adjustments, the account regression BLP performed better than the aggregate regression. Additionally, for the summer months, when events are more likely to be called, the adjusted aggregate regression performs the best.

Table 3: Aggregated Baseline Performance Accuracy Metrics Summary

Event Date	APE			
	Account Regression	Account Regression, Adjusted	Aggregate Regression	Aggregate Regression, Adjusted
18JAN2011	18. 2%	6. 0%	14. 3%	5. 7%
22FEB2011	8. 7%	3. 9%	16. 6%	6. 7%
31MAR2011	9. 8%	10. 4%	6. 5%	11. 6%
01APR2011	9. 9%	7. 0%	3. 7%	10. 6%
04MAY2011	8. 0%	9. 6%	12. 4%	10. 1%
27JUN2011	5. 7%	8. 1%	10. 5%	3. 8%
06JUL2011	14. 4%	9. 7%	10. 4%	3. 6%
26AUG2011	11. 1%	9. 8%	10. 8%	3. 9%
07SEP2011	11. 6%	10. 6%	14. 5%	3. 3%
13OCT2011	11. 4%	13. 9%	13. 7%	6. 9%
All	10. 7%	9. 3%	11. 3%	5. 9%

As with the accuracy metrics, the bias of the BLPs varied considerably depending on the specific proxy event. As seen in Table 4, there is a clear tendency for the adjusted aggregate regression BLPs to underestimate in the spring months. For the proxy event from March through June, a large majority of the BLPs underestimated the actual load, often by a large margin. The BLPs for July through September show far less bias, although there were some BLPs that highly overestimated or underestimated the load in August and September.

Table 4: Baseline Performance Bias Summary

Event Date	Account Regression		Account Regression, Adjusted		Aggregate Regression		Aggregate Regression, Adjusted	
	Median	Mean	Median	Mean	Median	Mean	Median	Mean
18JAN2011	**-18. 2%**	-18. 8%	**-6. 0%**	-5. 6%	**-14. 3%**	-14. 5%	**5. 7%**	4. 6%
22FEB2011	**-8. 7%**	-8. 2%	**-3. 4%**	-2. 7%	**-16. 6%**	-15. 7%	**-3. 6%**	-4. 4%
31MAR2011	**-9. 8%**	-11. 0%	**-0. 3%**	-2. 2%	**-0. 5%**	-1. 1%	**6. 3%**	4. 7%
01APR2011	**1. 7%**	1. 7%	**3. 9%**	6. 1%	**1. 8%**	4. 6%	**10. 6%**	10. 9%
04MAY2011	**3. 1%**	1. 2%	**9. 6%**	8. 1%	**12. 4%**	9. 2%	**10. 1%**	8. 4%
27JUN2011	**5. 4%**	4. 1%	**8. 1%**	7. 4%	**10. 5%**	10. 2%	**3. 8%**	5. 4%
06JUL2011	**14. 4%**	15. 6%	**9. 7%**	11. 9%	**10. 4%**	9. 6%	**-1. 3%**	0. 4%
26AUG2011	**11. 1%**	9. 3%	**6. 9%**	7. 6%	**10. 4%**	8. 2%	**0. 8%**	0. 1%
07SEP2011	**11. 6%**	13. 1%	**10. 6%**	10. 8%	**14. 5%**	14. 6%	**0. 2%**	0. 9%
13OCT2011	**11. 4%**	11. 2%	**13. 9%**	13. 7%	**13. 7%**	14. 2%	**6. 9%**	8. 7%
All	**5. 8%**	3. 5%	**7. 7%**	6. 5%	**8. 4%**	5. 8%	**3. 6%**	4. 2%

Summary and Conclusions

The analysis presented in this paper has demonstrated that estimating account-level BLPs for residential customers using traditional approaches presents substantial challenges in terms of achieving the desirable accuracy. The method – representative day vs. regression models – was not of great importance, as both produced BLPs with high levels of overall inaccuracy. As an alternative to account-level BLPs, this study also examined the effectiveness of estimating baselines using aggregated load. While the results were not conclusive, the aggregated model tended to produce more accurate BLPs. It is important to note that for both account-level and aggregate BLPs, the pre-event adjustments greatly improve the accuracy of the estimated loads.

The variability in residential load, with customers with inconsistent occupancy schedules, means that the estimated account-level BLPs will often result in extremely large prediction errors. There may be several consequences of using inaccurate settlement baselines. First, there will be a significant number of customers that get paid for not doing anything on event days and a significant number of customers will not get paid even though they did take action to lower their consumption during events. This could potentially create a customer satisfaction issue. Second, this situation can be viewed as creating a cost that does not result in a system benefit (i.e. paying significant incentives and getting no load reduction in return). Third, it also calls into question whether the load impacts for a default residential peak-time rebate program can be evaluated accurately.

The purpose of this analysis was to assess traditional settlement BLP methodologies. Given the issues with settlement baseline performance for these residential customers, some type of new approach to measuring and financially settling load reduction performance should be considered. For the representative day BLPs, this might involve exploring new adjustment methods or possibly moving to an individualized “firm-service” type of settlement model. Another possible area of research is in optimizing the period for calculating an adjustment factor. The two hours prior to the event period may not be the best performing. The most promising area for research is in the use of control groups to evaluate the impact of programs. This approach, however, may be problematic for default peak time rebate (PTR) programs where all eligible customers are participants.