Accuracy of Alternative Baseline Methods for Measuring Demand Response Program Impacts

Dr. Steven D. Braithwait, Christensen Associates Energy Consulting, Madison, WI

ABSTRACT

This paper addresses a continuing issue in the design of demand response (DR) programs – the performance of alternative baseline calculation methods, which are used to measure load reductions during DR events. The measured load reductions ultimately determine the amounts of DR credits paid to customers. The paper describes the results of two baseline analyses conducted using data from DR programs operated by the major investor-owned utilities in California. The performance of a range of alternative baseline methods was examined with regards to both *accuracy* and *bias* (*e.g.,* the tendency of a baseline method to under-state or over-state the "true" baseline). Data were used for both actual events, as well as *event-type* days of similar high temperatures and system load conditions (or actual events for customers who did not actually participate in those events). The use of event-type days was designed to provide cases in which the "true" baseline is represented by customers' *observed* usage during a "pseudo-event" period. Overall results, which were somewhat mixed, suggest that baseline performance depends to a large extent on customer characteristics (*e.g.*, business type, degree of weather sensitivity, and inherent load variability), and on the nature of event days and the days that make up the representative days for constructing the baselines. However, certain patterns were evident. First, all of the baseline methods applied to *commercial*-type customer accounts tended to be more accurate and less biased than for *industrial*-type or school accounts. Second, the *adjusted* 10-in-10 baseline tended to perform best overall.

1. Introduction

The utility industry in the United States has taken some steps toward restructuring, particularly on the wholesale side, with organized wholesale energy markets operated by Independent System Operators (ISOs), and private organizations allowed to build generation and sell it into these wholesale markets. Several of these wholesale power markets have experienced occasional price spikes (*e.g.*, energy prices that have on occasion exceeded $1,000/MWh) under conditions of unusually high demand. It has been recognized that the lack of price-responsive demand on the retail side has been a primary cause of these price spikes. That is, since nearly all electricity consumers face fixed retail energy prices, they have no incentive to reduce consumption during periods of high wholesale prices since those prices are not passed on in retail rates. Given this fact, most of the regional ISOs (as well as some utilities) have established demand response (DR) programs, in which consumers, typically acting through third-party curtailment service providers, or aggregators, are able to bid *load reductions* into day-ahead, hour-ahead, or real-time energy markets on occasions when prices reach a certain level, and receive payments for the amount of the load reduction. A key element of any DR program is the mechanism used to calculate customers' *baseline load*, or an estimate of the hourly usage level that the customers would have consumed in the absence of the DR event.

This paper addresses a continuing issue in the design of DR programs, which involves the performance of various alternative baseline calculation methods. Baselines are used as the basis for estimating load reductions during DR events, and thus determine the amount of payments to DR participants. A method for estimating baselines is needed because load *reductions* cannot be measured

directly; only energy that is *consumed* can be metered. A baseline load method is therefore needed to construct the counter-factual load that would have occurred had the event not occurred.[1]

A wide variety of baseline methods have been used by utilities and ISOs. Some are as simple as setting the baseline equal to the customer's energy consumption in the hour prior to the event. This approach can work reasonably well for emergency DR programs that are called at short notice and last for a short time. However, non-emergency programs are often called with day-ahead notice. Baseline load methods for these types of programs generally involve same-hour averaging across a certain number of days prior to an event. Typical features of baseline methods involve how many days are included in the average, whether the days selected should meet any criteria (*e.g.*, the three days out of the previous ten days that had the highest energy usage during the typical event hours), and whether the resulting day-averaged baseline load should be adjusted using pre-event data on the day of the event (*e.g.*, in an attempt to adjust for the fact that the event day may be characterized by more severe weather conditions than those on the days included in the average, and thus cause customers' loads to be correspondingly higher).

This paper reports on two baseline analyses that we conducted for the major investor-owned California utilities as part of load impact evaluations of the DR programs involved[2]. In one program, third-party aggregators enroll lists of customers and serve as intermediaries between the customers and utilities in arranging load reductions during DR events. In this case, the baseline applied to aggregations of participating customers rather than individual customers. In another program, customers sign up directly to participate in a demand bidding program (DBP), and have the opportunity to submit bids for load reductions on days for which the utility "calls" a DBP event.[3] In this case, individual customer baselines were analyzed. Customers are paid for load reductions, though performance is optional, with no penalties for non-performance. For the aggregator program, we used data for both events called by the utilities, as well as a selection of ten non-event, but *event-type* days of similar high temperatures and system load conditions. The use of event-type days was designed to provide cases in which the "true" baseline is known and represented by customers' or aggregators' *observed* usage during a "pseudo-event" period. For the event days, the customers' baselines cannot be observed because they may have reduced usage during the event. We assumed that the most accurate estimate of their baseline load was provided by a regression analysis of their load during the entire summer period, and thus constructed implied baseline loads from customer-specific regression models which were based on data for that period. For the DBP program, we only compared alternative baselines to a regression-based baseline for event days.

2. Alternative Baseline Load Methods

The alternative baseline load methods that were tested were those in which the utilities were most interested, including the method that has been used for the last several years (known as the 3-in-10 method), and the method that has been proposed for the future. All involved day-averaging over all or subsets of the ten most recent non-event weekdays, along with possible adjustments to the resulting baseline. The alternatives are as follows:

- *10-in-10 baseline*. Select the most recent 10 weekdays prior to the event day, where these excluded weekends, holidays, and any other previous event days. Then calculate the baseline load for each hour of the DR event period (curtailment period) as the average for that hour over the 10 days.

---

[1] Estimating the baseline load for DR programs is analogous to estimating the amount of energy that energy efficiency (EE) program participants would have consumed had they not participated in the program, which is a fundamental element of EE program impact evaluations.

[2] See [CAEC 2009] and [CAEC 2010]. Annual load impact evaluations are mandated by the California regulators to estimate the hourly load reductions for each event that was called in that year, for each DR program.

[3] Utilities typically call DR events when temperatures exceed a certain level or the utility's system load is expected to exceed a predetermined level.

- *5-in-10 baseline*. Identify the 5 days out of the 10 that had the highest "overall consumption during the curtailment hours." Then calculate the baseline load for each hour of the curtailment period as the average for that hour over the 5 selected days.
- *3--in-10 baseline*. Calculate the baseline load for each hour of the curtailment period as the average for that hour over the 3 days with the highest curtailment-period usage.
- *Adjusted baselines*. Adjust the 3, 5 and 10-in-10 baseline methods by applying a scalar adjustment to the unadjusted baseline, where the adjustment takes the form of the ratio of the average hourly usage during specified pre-event (*i.e.*, morning) hours on the event day to the usage in the same period for the unadjusted baseline load. In our analysis, average usage for the four hours-ending 8 through 11, prior to the event period, was used for calculating the scalar adjustments.

The adjusted baseline potentially provides improved accuracy, particularly for weather-sensitive customers whose pre-event usage is not affected by actions taken in response to the event (*e.g.*, increases in usage due to pre-cooling, or reductions in usage due to early shut down of production processes). However, there is a risk that for customers who do take such actions, an event-day adjustment could actually *increase* the error of the unadjusted baseline.

3. Baseline Performance Statistics

Previous analyses[4] of the accuracy of alternative representative-day, or day-averaging baseline load methods have focused on certain statistical measures designed to measure factors such as *accuracy* (average or median differences between the estimated baseline load and the actual, or simulated load); *bias* (whether the estimated baseline load is systematically higher or lower than the actual); and *variability*, or the extent to which errors for some customer types tend to be larger than for other types.[5]

In this study, baseline **accuracy** was measured using the *relative root mean square error* statistic (RRMSE). This statistic measures the relative degree of difference, or error, regardless of sign, between two data series, which in these studies are the alternative baselines and either the actual baseline (for customers who did not participate in a particular event) or the regression-based baseline (as the presumed best estimate of the true baseline). This statistic is nominally bounded by 0 and 1, with values closer to 0 indicating greater accuracy. Since the root-mean squared *errors* are normalized by the root-mean squared *load levels*, the resulting statistic is a normalized, or percentage measure of accuracy relative to the true baseline. For example, a value of 5 percent indicates an average 5 percent error in the baseline load (or difference between an alternative program baseline and the regression-based baseline) relative to its mean value.

The formula for this statistic is the following:

$$RRMSE = [(1/n) \sum (e_h)^2]^{1/2} / [(1/n) \sum (L^A_h)^2]^{1/2},$$

where in this case

e_h $= (L^A_h - L^P_h)$,
L^A_h is the actual or regression-based baseline load,
L^P_h is one of the alternative *predicted* (program) baseline loads,
n is the total number of customer event days and hours, and the sum is across event days and hours, for each sub-group of customers (*e.g.*, by industry type).

[4] See [KEMA 2003], [Quantum 2004], [Quantum 2006], and [LBNL 2007] for previous baseline analyses.
[5] Baseline accuracy is important for understanding the confidence that may be placed in DR as a substitute for costly peaking generation, and for justifying payments to customers for the load reductions. Bias is important for assuring that customers are not systematically under-compensated or over-compensated for their load reductions. The degree of variability of estimated baselines provides guidance on types of customers for whom accurate baselines are unlikely and who may be best directed to dynamic pricing tariffs for which baseline calculations are not necessary.

Median values of RRMSE statistic. The relative errors of the baseline estimates can reach extremely high levels for accounts that have a number of observations with very small actual loads. These large errors can distort any "average" measure of error, or accuracy *across* customer accounts. For that reason, the *median* value of the statistics is a useful indicator of the typical accuracy of each baseline method across accounts of a particular type.

Bias was measured using the *median percent error*, where the percent error is defined as the *difference* between the "true" baseline load (or the regression-based baseline for event participants) and an alternative estimate of the baseline load, divided by the *level* of the true baseline. Using this convention, positive errors indicate *downward bias* (*i.e.*, the true baseline exceeds the estimated baseline), and negative errors indicate *upward bias* (*i.e.*, the estimated baseline exceeds the true baseline).

The median percent error statistic is the median value of all of the percent errors calculated across customers and event hours, for a particular customer type (*e.g.*, commercial or industrial). This statistic indicates the extent to which a given baseline method tends to *over-state* or *under-state* the true baseline. While the median statistic serves to indicate the *typical* bias tendency, we have found that examining the *distribution* of percent errors provides greater insight into the larger pattern of differences in the alternative baselines. Thus, we show actual distributions of all percent errors, or *deciles* of the distribution of percent errors (where the value that determines the 50^{th} percentile is the median value of the distribution).

4. Results for Aggregator Event-type Days

This section reports baseline performance statistics for both unadjusted versions of the baseline methods described above, and adjusted versions that use information on energy consumption in pre-event hours on event days to adjust the baseline upward or downward. Results are shown for three broad industry types – *commercial* (including offices, retail stores, hotels and government administration), *industrial* (including manufacturing, wholesale, water utilities and pipelines), and schools.

4.1 Accuracy of unadjusted baselines

Figure 1 shows *accuracy* results for unadjusted versions of the three alternative baseline methods that use different numbers of days in the baseline calculation (*e.g.*, 3, 5, or 10). The bars in the figure are grouped first by aggregator, then by industry type. The three bars in each panel show results for the different baseline calculation methods (*e.g.*, 3, 5, or 10-in-10). The following observations characterize some of the notable results:

- For the unadjusted 3-in-10 baseline, and focusing first on the sets of columns labeled TOTAL for each aggregator, median relative errors range from about 5 to 6.5 percent across the aggregators, with an overall relative error of 5.6 percent.
- Moving to the 5-in-10 and 10-in-10 bars, the overall relative errors are generally higher than for the 3-in-10.[6] Patterns across aggregators and industry types vary.
- Comparing results by industry type, the findings suggest that the relative errors in unadjusted baselines for *commercial* customers are generally larger than those for *industrial* customers, and frequently are higher when more days are included in the baseline calculation.
- For both methods, *schools* often have among the highest relative errors.

[6] This result is not unexpected, particularly for commercial customers whose usage is weather sensitive, and for which usage on event days may be better represented by the three days of highest usage out of the previous ten.

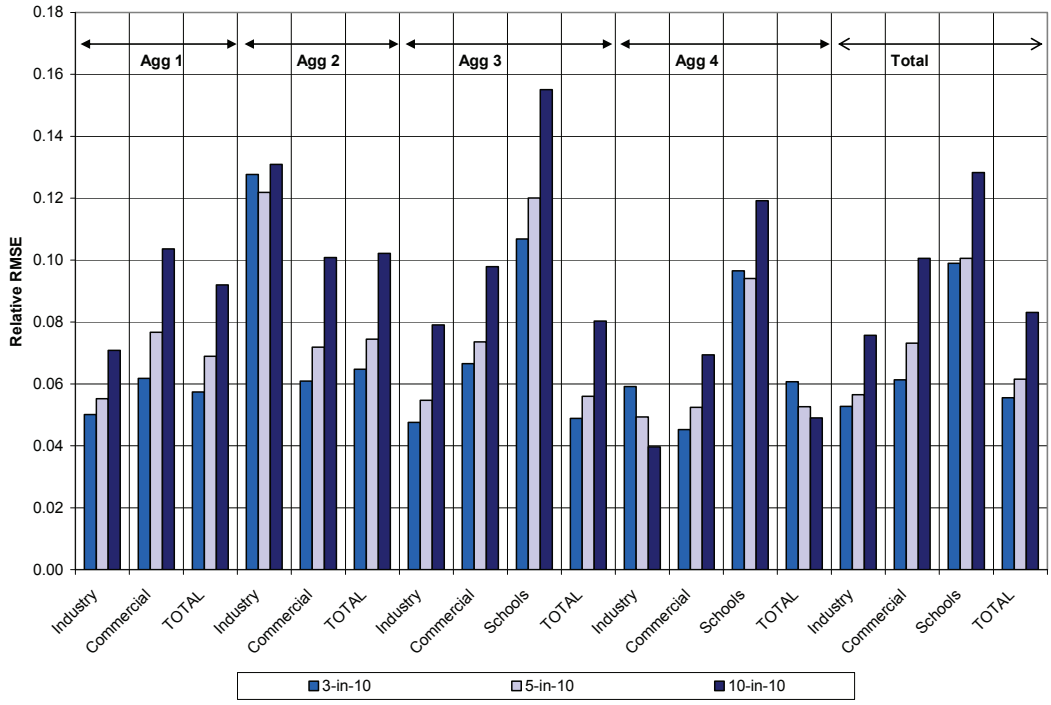

Figure 1. Accuracy of *Unadjusted* Baselines

4.2 Bias of unadjusted baselines

Figure 2 presents comparable results for the *bias* of the unadjusted baselines, showing the median percent errors across event-type days and hours, both by aggregator and overall. As noted above, *positive* errors (*i.e.*, estimated baseline is less than actual) indicate *under-stated* baselines, or downward bias, and *negative* errors indicate *over-stated* baselines, or upward bias. Observations include the following:

- The values shown in the TOTAL bars are positive for each aggregator and overall, indicating that the unadjusted 3-in-10 aggregator baseline is biased downward (*i.e.*, typically *under-states* the true baseline), by about 2 to 3 percent.
- The overall downward bias of the unadjusted baseline tends to grow larger as the number of days included in the baseline average increases. This is not unexpected, particularly for weather-sensitive customers, as the additional included days may be increasingly milder than the event-type days.
- Looking at industry types, the downward bias of the unadjusted baselines is generally larger for *commercial* than for *industrial* customers (for aggregator 4, the 3 and 5-in-10 show *upward* bias), and the difference is greater with more days included in the baseline.

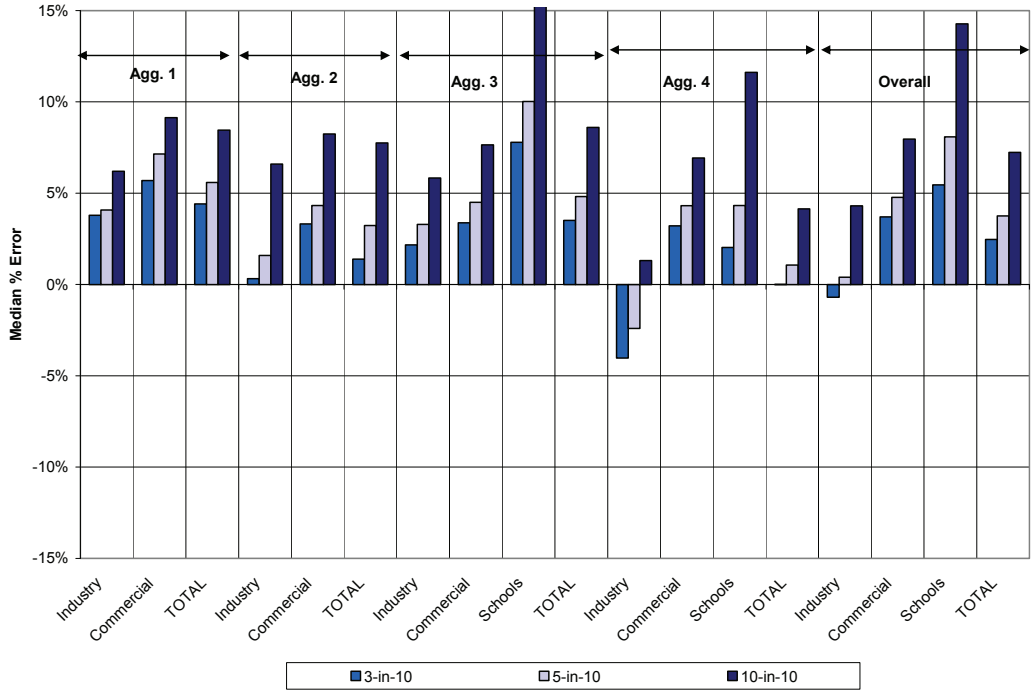

Figure 2. Bias of *Unadjusted* Baselines

4.3 Accuracy of Adjusted Baselines

Figure 3 reports accuracy results for the alternative adjusted baseline methods. Key findings include the following:

- Focusing first on the TOTAL bars, the adjustment generally improves baseline accuracy substantially, reducing relative errors by half or more in many cases compared to the unadjusted baselines (compare to Figure 1).
- The adjusted 5-in-10 and 10-in-10 baselines are substantially more accurate than the unadjusted, with relative errors approximately half that of unadjusted versions.
- Looking across industry types, the adjusted baselines for *commercial* customers are generally more accurate than those for *industrial* customers, though in many cases the differences are not great, and the adjusted baselines for *schools* are the least accurate.

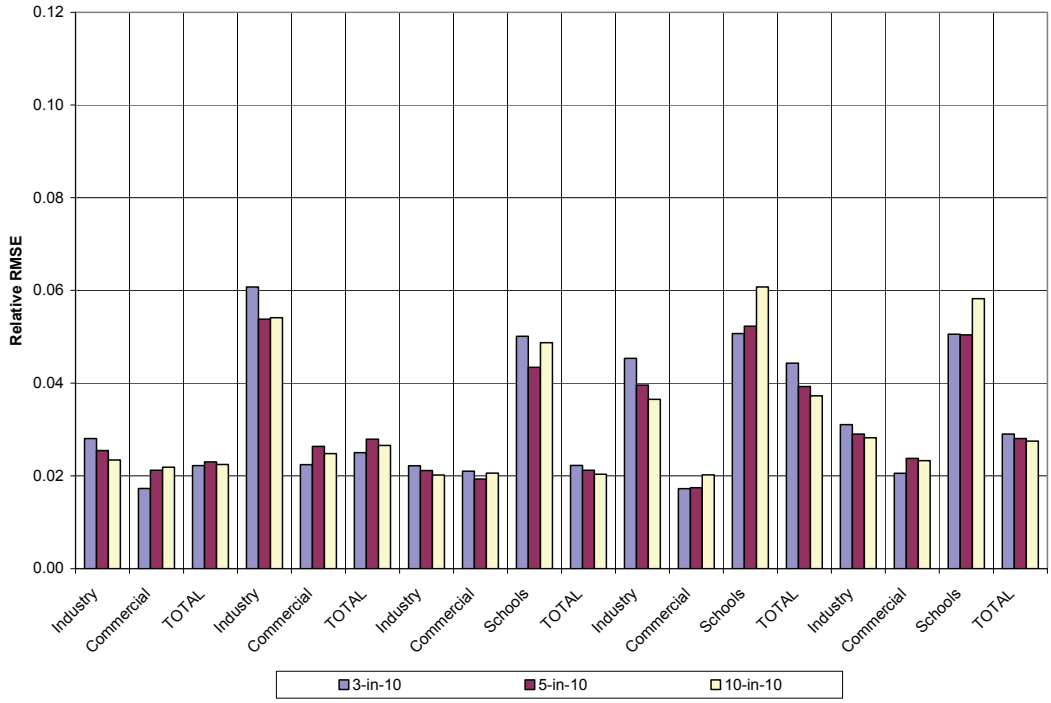

Figure 3. Accuracy of *Adjusted* Baselines

4.4 Bias of adjusted baselines

Figure 4 reports bias results for the alternative adjustment methods. Key results are the following:

- At the TOTAL level, the morning adjustments generally reduce the magnitude of the bias, typically converting a downward bias (under-statement) of the unadjusted 3-in-10 baselines to a small upward bias (*e.g.*, a negative value of less than one percent for three of the four aggregators).
- Looking across columns as the number of days included in the *aggregator* baseline increases, the extent of the small upward bias appears to decrease, to the point that the biases for the adjusted 10-in-10 baseline are very small under-statements or over-statements. Across all customers, the median % error is less than one percent.
- Looking *across industry types*, there are few consistent patterns for the aggregator baselines, though the biases for commercial types are generally smaller than industrial, and the adjusted 10-in-10 baseline usually shows the smallest bias.

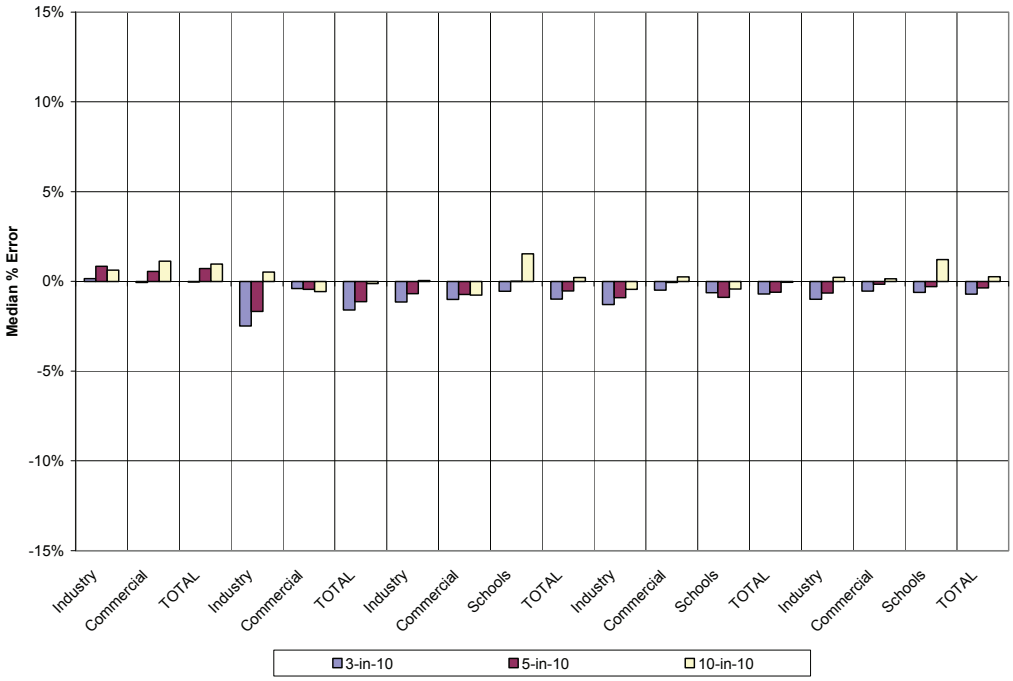

Figure 4. Bias of *Adjusted* Baselines

4.5 Customer-level distributions of baseline errors

Figure 5 illustrates the variability in relative errors at the customer level, which underlie the distributions shown in the above figures. It shows the distributions of percent errors for unadjusted and adjusted 10-in-10 baselines. The points, which represent the average percent error for a customer and event-type day, are sorted by the values for the *unadjusted* baselines, thus providing an indication of the improvements in the percent errors due to the morning adjustments, as well as the breadth of the distributions across customers and event-type days. The unadjusted baseline *under-states* the true baseline in more than two-thirds of the cases (*i.e.*, the curve crosses the horizontal axis less than a third of the way from the origin), which is consistent with an estimated median percent error of positive 7.2 percent.[7] The *adjusted* baseline points show a relatively high density within about 5 percent on either side of the horizontal axis (see 5% lines in the figure), thus indicating the extent to which the adjustments reduce the baseline errors. The median percent error for the adjusted baseline across all customers and event-type days is 0.81 percent. The bounds on the distribution of errors for the adjusted baseline are due to the 80% limits set on the morning adjustment.

[7] Very large baseline over-statements (the initial tail of the distribution) occur when a customer's actual load during the event period on an event-type day is quite low relative to a baseline calculated by averaging usage across several previous days of irregular loads (*e.g.*, 100 kW actual load compared to a baseline load of 500 kW), resulting in a large negative error divided by a small actual baseline, thus producing a very large negative value (*e.g.*, (100 – 500) = –400, divided by 100, which implies a relative error of –400 percent). Recall that this baseline analysis used event-type days on which the customers did not actually face an event, and thus had no incentive (other than the existing peak demand charge) to reduce load.

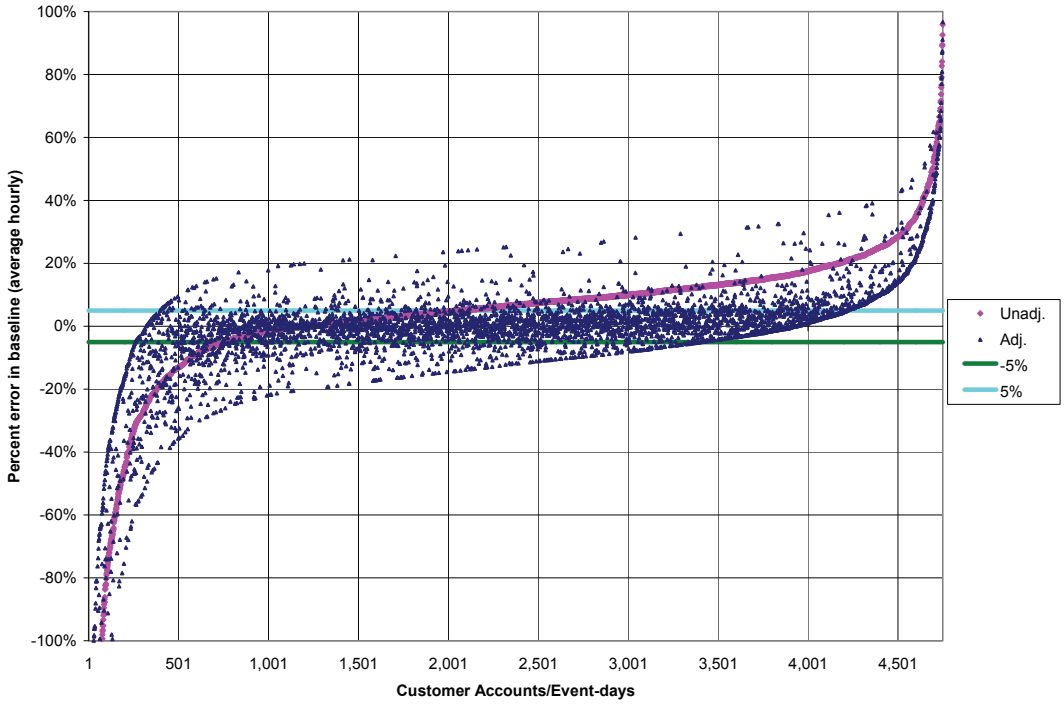

Figure 5. Distributions of Average Percent Errors for *Unadjusted* and *Adjusted 10-in-10* Baselines – *Individual Customer Accounts*

4.6 Conclusions of aggregator baseline analysis

The results of the aggregator baseline analysis using event-type days provide a reasonably consistent story regarding the baseline issues of the relative accuracy of aggregator baselines, and the effect of morning adjustments to 3-, 5-, and 10-in-10 baselines on the bias of unadjusted baselines. Some results are mixed, suggesting that baseline performance depends on the characteristics of customers and event days. Major findings include the following:

1. Regarding the effect of *morning adjustments* to the 3-in-10 baseline on *bias*, the results suggest that the adjustments do improve the bias of the unadjusted baseline relative to the "true" baseline.
2. Expanding the analysis to consider adjusted 5-in-10 and 10-in-10 baselines produced results suggesting that the *adjusted 10-in-10 method* may produce both the greatest accuracy and the smallest bias.
3. Comparing unadjusted 5-in-10 and 10-in-10 baselines to comparable adjusted versions illustrates the improved performance of the adjusted versions.
4. Examination of the variability of percent errors of 10-in-10 baselines across the *individual customers* that made up the aggregator loads illustrates the source of baseline errors at the aggregator level.

The performance of the alternative baseline methods on event days, in terms of accuracy and bias, was qualitatively similar to their performance on the *event-type* days presented in Section 4. In particular, adjusting the baseline for morning usage generally improves the accuracy and reduces the bias of the unadjusted baselines.

5. Demand Bidding Baseline analysis

In a separate baseline analysis, using data from the demand bidding program offered by Southern California Edison (SCE) in 2009, we compared the current program 3-in-10 baseline method to the

unadjusted and adjusted 10-in-10 baseline (which is scheduled to replace the unadjusted 3-in-10 in future years), and the baseline implied by the econometric estimates of load impacts developed in the load impact evaluation for the 2009 program year (impact evaluations are undertaken annually). We used customer-level load data to calculate event-day baseline loads for those DBP participants who submitted bids to calculate baselines using the following methods:

- The 3-in-10 method currently used in the program;
- The 10-in-10 method, unadjusted for pre-event load levels;
- The 10-in-10 method with an adjustment for pre-event load levels, where the adjustment factor takes the form of the ratio of the average hourly usage in the four hours prior to the event to the average over the same hours from the 10 weekdays from which the 10-in-10 baseline is calculated, and the adjustment is limited to no more than 20 percent.[8]

Since we wished to assess alternative baselines for days on which events actually occurred, we could not use customers' observed loads as the "true" baseline for comparison purposes. Our assumption was that the most accurate estimate of the true baseline was that provided by a regression analysis of customers' hourly loads for the entire summer period. The baseline implied by the regression model for a particular customer was derived by adding the regression's estimates of hourly load impacts during each event to that customer's *observed load* during the event hours.[9] This produced an estimate of what the customers' usage would have been had the event not occurred. For example, if a customer's observed load during an event was 800 kW in each hour, and the estimated load impact was 200 kW in each hour of the event, then the implied baseline load would be the sum of the two values, or 1,000 kW per hour. That baseline load then served as the standard to which the alternative program baseline loads were compared.

To examine potential differences in baseline performance by customer type, customers were classified into one of three business-type categories, as described in Section 4 – *Industrial* customers, who were assumed to be not particularly weather sensitive; *Commercial* customers, who were presumed to be weather sensitive; and *Schools*, whose load patterns often vary considerably during summer months due to vacation schedules for which information is often not available, and who were therefore treated as a separate industry type. Percent errors were calculated for each of the eight hours (HE 13 – 20) of each of the 14 events in which a customer submitted a bid, and thus may be expected to have reduced load during the event. Approximately 500 customer accounts submitted bids for at least one event.

5.1 Accuracy

In the case of the DBP analysis, results are shown for all three alternative baselines together, including the adjusted 10-in-10. Figure 6 summarizes the *accuracy* results for the alternative baselines compared to the regression-based baseline for SCE's DBP bidders, for each of the industry groups. The results indicate that for the large number of industrial-type customer accounts, in the case of the 2009 events, the program baselines (based on the unadjusted 3-in-10 method) were least accurate compared to the regression-based baseline. Median differences ranged from over 30 percent for the unadjusted 3-in-10 baseline to 25 percent for the *adjusted* 10-in-10 baseline. Differences between baselines were smallest for the commercial-type customer accounts, ranging from 7.4 percent for the unadjusted 3-in-10 to 4.4 percent for the adjusted 10-in-10 baseline. These results are not too surprising, since many industrial customers' loads tend vary more widely from day to day for unknown reasons than do commercial customers' loads, which can reduce the

[8] These features of the adjusted baseline have been negotiated between the utilities, regulators and various customer organizations.

[9] Except for regression errors, this calculation is equivalent to simulating the load on the event day using the estimated regression coefficients, with values for all explanatory variables inserted for the event day, but with the event variables "turned off" (*i.e.*, set equal to zero).

accuracy of any baseline method.

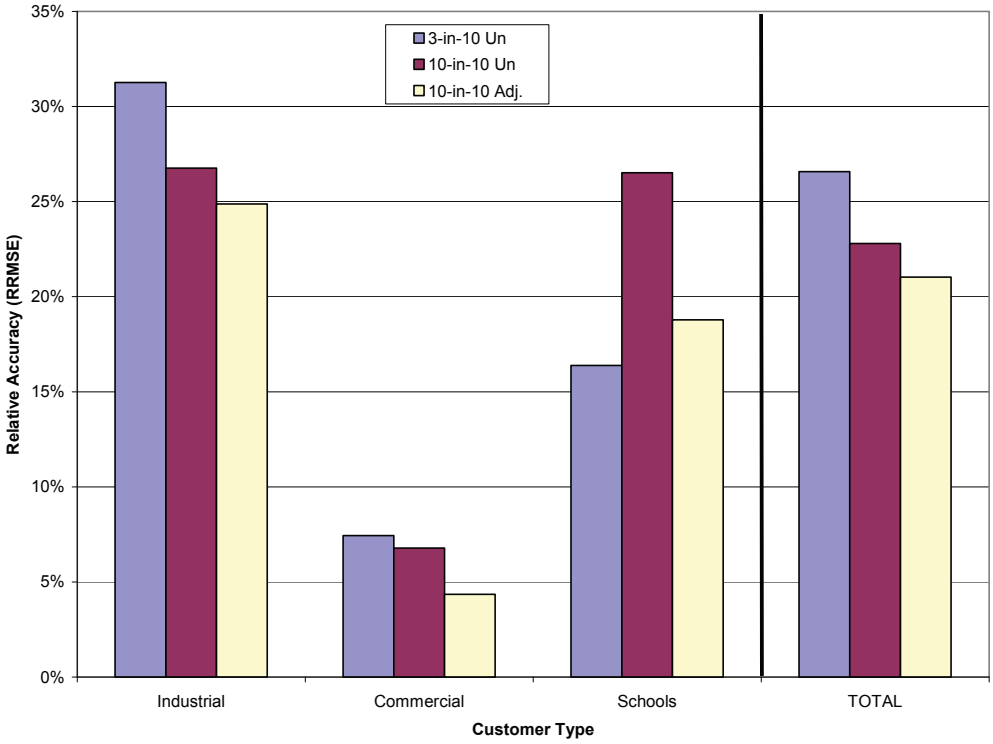

Figure 6: Accuracy of Alternative Baselines – SCE DBP (*Median RRMSE*)

5.2 Bias

Figure 7 presents results for the typical *bias* of the alternative baselines relative to the regression-based baseline. The unadjusted 3-in-10 results suggest that the current program baseline typically *over-states* load impacts for both the industrial and commercial categories (negative values), while *under-stating* load impacts for school customers (relative to the regression-based estimate). The results for the adjusted 10-in-10 baselines indicate a reduction in the typical biases for industrial and commercial accounts, but not for schools, whose typical 6 percent under-statement grows to nearly 12 percent. Additional insight into the range of baseline errors across customer accounts is provided in the figures below.

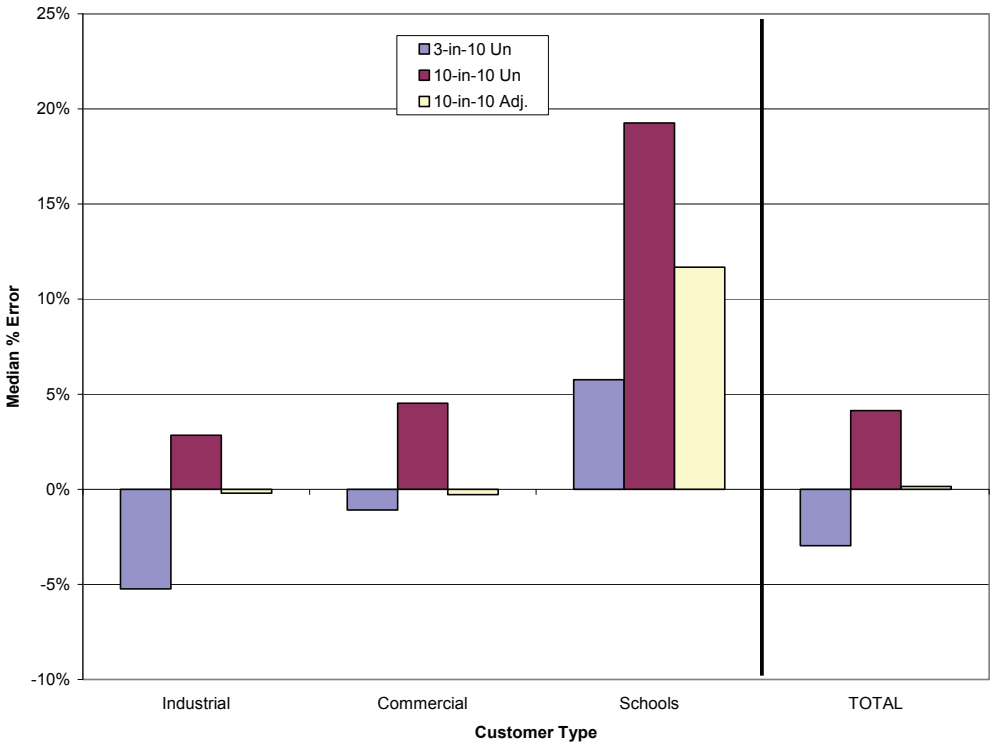

Figure 7: Bias of Alternative Baselines – SCE DBP (*Median Percent Error*)

5.3 Distribution of percent errors

The median of the distribution of percent errors provides a useful summary measure of the typical bias of a particular baseline method. However, information on the nature of the full distribution provides greater insight into the nature of the percentage baseline errors across customers. Figure 8 expands on the single median value of the percent errors between the three alternative baselines and the regression-based values by presenting values that determine *deciles* of the range of percent errors. Nine values are provided for each baseline method and customer type, each representing boundary values that separate 10 percent of the customer-hour values ordered by size. The 50 percentile values represent the median values of the distributions of errors reported above. Thus, for example, the median percent error for the unadjusted 3-in-10 baseline for the industrial-type customers is negative 5.2 percent, indicating a modest "typical" over-statement relative to the regression-based baseline. However, the 30th percentile value indicates that 30 percent of the over-statements exceed 16 percent, while the 70th percentile value indicates that another thirty percent of the values reflect *under-statements,* but that are relatively small, exceeding only 0.5 percent. The distributions for the commercial-type customer accounts are generally "tighter," with fewer large percent errors.

Three features of the distributions of percent errors for the alternative baselines stand out. First, for all three industry types the decile values for the 3-in-10 baselines tend more toward the negative direction (*i.e.*, to be more negative or less positive) than the 10-in-10 baselines. Again, this makes sense, as the 3-in-10 baseline is averaged over the three highest loads in the 10-in-10 baseline, and thus should always be at least as large as that baseline. Second, for the *commercial* customers, between 40 and 50 percent of the 3-in-10 values and more than 70 percent of the 10-in-10 values are positive, indicating *under-statements* relative to the regression-based baseline. Third, for both the industrial and commercial customer accounts, the *adjusted* 10-in-10 baseline generally reduces the percent errors (compared to the unadjusted 10-in-10) and shifts the distribution of percent errors toward the origin (*i.e.*, zero error). In addition, the large values at the negative end of the distribution for industrial customers suggest fairly large baseline errors for a number of DBP bidders of that type.

Finally, the distributions of percent errors for *schools* suggest that all of the alternative baselines tend to under-state the baselines as measured by the regression equations for at least 70 to 80 percent of

the customer-hour observations (*i.e.* all but the 10th, 20th or 30th decile values are positive). These results again serve to indicate the frequent difficulty of determining appropriate baseline loads for schools during the summer months.

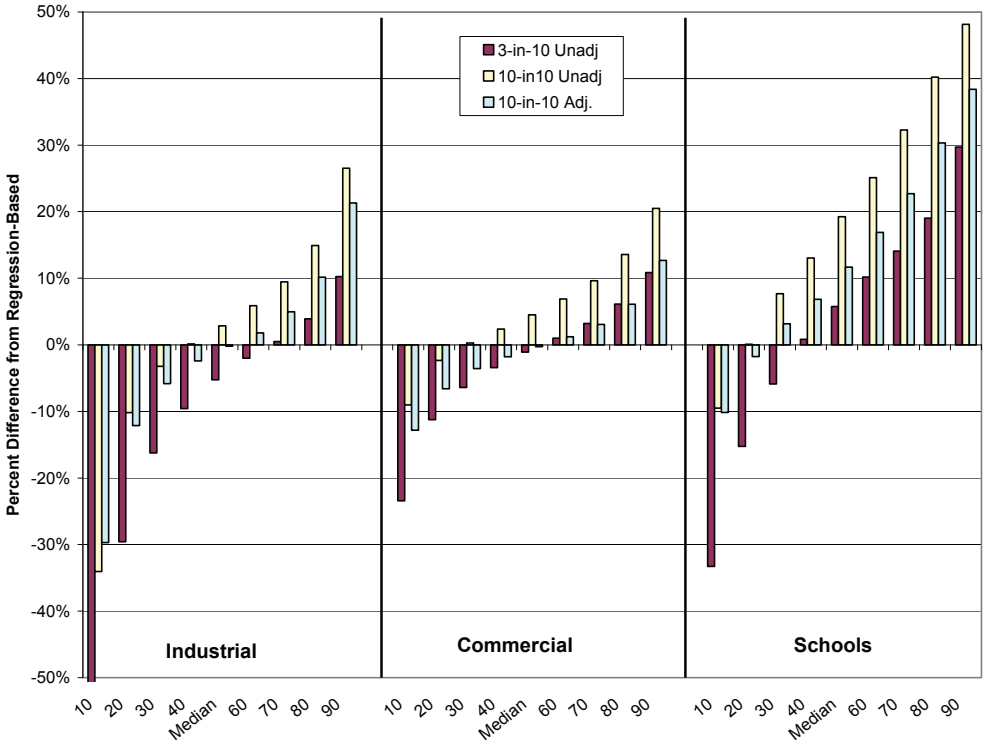

Figure 8: Percentiles of Percent Errors of Alternative Baselines – *SCE DBP*

5.4 Summary conclusions

The comparison of alternative baseline methods for the DBP customer accounts points to several consistent findings. First, all of the baseline methods applied to *commercial*-type customer accounts tended to be more accurate and less biased relative to the regression-based baseline than they did for industrial-type or school accounts. Second, the unadjusted 3-in-10 program baseline tended to over-state the regression-based baseline by more than did the unadjusted 10-in-10 baseline (which is not surprising since the 3-in-10 uses the 3 days with highest loads from among the 10 available). Third, the *adjusted* 10-in-10 baseline tended to reduce both over-statements and under-statements of the unadjusted baseline, and would thus be likely to improve accuracy and reduce bias in calculating load impacts for DBP, compared to unadjusted versions of either the 3-in-10 or 10-in-10 baseline.

6. Conclusions and Recommendations

The baseline performance results for both the aggregator loads and individual customer loads in the demand bidding program illustrate several key findings and suggest certain recommendations regarding baseline methods. Perhaps most importantly, both sets of results indicate that adjusting baselines that are constructed as averages across recent days, using pre-event usage data on the day of an event seems to improve the accuracy and reduce the bias of unadjusted baselines. The adjusted 10-in-10 method appears to produce the smallest baseline errors. The adjustment method essentially establishes the *shape* of the hourly baseline load by averaging usage across recent days, and then adjusts the *level* of the baseline to the morning usage on the event day. For weather-sensitive commercial customers, the adjustment can raise the baseline to reflect higher than normal usage on event days, which typically occur on days that are hotter than normal. For industrial customers, which tend to be much less weather

sensitive than commercial customers, but have more variable loads, the adjustment appears to bring the baseline closer to the usage level that would otherwise have occurred had an event not been called.

While an adjusted baseline can reduce the *typical* baseline error to the range of five percent or less, examination of the full distribution of percent errors across all customers indicates that relatively large baseline errors may occur for a substantial share of the participating customers (*e.g.*, 20 percent of customers may have baselines that over-state or under-state their true baseline by 20 percent or more). This result suggests two possible recommendations. One is to establish limits on the load variability of customers that wish to participate in DR bidding programs that require baseline measurement.[10] The other is to re-design DR programs such that baseline load levels on event days are pre-established contractually, as in a forward contract, and event-day load reductions are calculated relative to the pre-established baseline.[11]

References

[CAEC 2009] Steven D. Braithwait and David Armstrong, Christensen Associates Energy Consulting, *2008 Evaluation of California Statewide Aggregator Demand Response Programs – Volume 2: Baseline Analysis of AMP Demand Response Program*, March 30, 2009.

[CAEC 2010] Steven D. Braithwait, Daniel G. Hansen, and Jess D. Reaser, Christensen Associates Energy Consulting, *2009 Load Impact Evaluation of California Statewide Demand Bidding Programs (DBP) for Non-Residential Customers: Ex Post and Ex Ante Report*, April 1, 2010.

[LBL 2008] Coughlin, K., M.A. Piette, C. Goldman, and S. Kiliccote, *Estimating Demand Response Load Impacts: Evaluation of Baseline Load Models for Non-Residential Buildings in California*, Demand Response Research Center, LBNL-58939, 2008.

[KEMA 2003] Goldberg, M.L. and G.K. Agnew, *Protocol Development for Demand-Response calculations: Findings and Recommendations*, Prepared for the California Energy Commission by KEMA-Xenergy, CEC 400-02-017F, 2003.

[Quantum 2004] Quantum Consulting and Summit Blue Consulting, *Working Group 2 Demand Response Program Evaluation – Program Year 2004 Final Report*, Prepared for the Working Group 2 Measurement and Evaluation Committee, 2004.

[Quantum 2006] Quantum Consulting and Summit Blue Consulting, *Evaluation of 2005 Statewide Large Nonresidential Day-ahead and Reliability Demand Response Programs*, Prepared for Southern California Edison and the Working Group 2 Measurement and Evaluation Committee, 2006.

[10] Where advanced metering has been installed, as in California, historical hourly interval usage data for all customers are available, which would allow calculation of various statistics on the variability of customers' hourly usage pattern across days and months.

[11] This approach, which is described in Chao (2010), would operate similarly to existing two-part real-time pricing programs, in that consumers would pay a fixed retail rate for the amount of the pre-established baseline load on event days, and receive credits at the wholesale price (or DR credit level) for any load reductions below the baseline level (customers would pay for any usage above the baseline during event hours at the same wholesale price, which would provide an incentive to reduce consumption).