2008 Evaluation of California Statewide Aggregator Demand Response Programs

Volume 2: Baseline Analysis of AMP Demand Response Program

Steven D. Braithwait and David Armstrong

Christensen Associates Energy Consulting, LLC

4610 University Ave., Suite 700
Madison, WI 53705
(608) 231-2266

March 30, 2009

Table of Contents

List of Tables

List of Figures

Abstract

This report summarizes the results of a baseline analysis that was undertaken as part of the statewide ex-post evaluation of the Aggregator demand response programs. The objective of the baseline analysis was to assess the relative accuracy and bias of several alternative methods for calculating baselines for measuring load impacts for settlement. Data for some 600 customers enrolled in PG&E's Aggregator Managed Portfolio (AMP) program were used to assess the performance of a range of unadjusted and adjusted baseline methods. Of particular interest were differences in performance between baselines calculated for the aggregation of load data across customers enrolled for a particular aggregator, and a baseline calculated as the sum of individual customers' baselines.

Executive Summary

This volume documents the results of a baseline analysis study undertaking in the context of an evaluation of aggregator demand response ("DR") programs operated by the three California investor-owned utilities (IOUs), Pacific Gas and Electric ("PG&E"), Southern California Edison ("SCE"), and San Diego Gas and Electric ("SDG&E") for Program Year 2008. In these programs, aggregators contract with commercial and industrial customers to act on their behalf to arrange load curtailments, receive incentive payments, and pay penalties (if warranted) to the utility. Each aggregator forms a "portfolio" of individual customers such that their aggregated load participates in the DR programs.

This study addressed a continuing issue in the design of such DR programs, which is the accuracy and bias of various alternative baseline methods that might be used to calculate the *baseline load* that is used to measure load reductions during events. Of particular interest for the aggregator programs are four issues:

1. Whether baselines should be constructed using the aggregated load of all enrolled customers who are nominated by the aggregator for the month in which an event is called, or by calculating baselines for each such customer, and summing the results.
2. How many days prior to an event should be included in the baseline calculation (*e.g.*, the three, five, or 10 days with the highest event-period consumption in the previous 10 days).
3. Should the baseline be adjusted using event-day usage data in an attempt to avoid understating an aggregator's "true" baseline on event days that tend to have more severe weather than the days prior to the events?
4. Was gaming avoided for the customers and aggregators who selected an adjusted baseline option in PG&E's Aggregator Managed Portfolio ("AMP") program in 2008?

The study used data for all AMP customers who were nominated during the summer months. The performance of a range of alternative baseline methods in terms of accuracy and bias (*e.g.,* the tendency of a baseline method to under-state or over-state the true baseline) was examined using data for both the five AMP events in 2008 and a selection of ten event-type days of similar high temperatures and PG&E system load. For the event-type days, the customers' or aggregators' observed loads during a pseudo-event period served as their true baseline. For the event days, the estimated hourly load reductions from the ex post load impact evaluation for each customer were added back to their observed load to create the true baselines.

To examine potential differences in baseline performance by type of customers, each aggregator's customers were classified into three categories of industrial, commercial and schools, and aggregate loads for those sub-groups were calculated for each aggregator.

Baseline performance was measured primarily by two statistics that have been used in previous baseline studies. Accuracy was measured using the *relative root mean square error* statistic (RRMSE, sometimes referred to as the Theil U-statistic). Bias was measured using the *median percent error*, where positive errors indicate downward bias, and negative errors indicate upward bias.

The results of this baseline analysis provide a reasonably consistent story regarding the baseline issues of the relative accuracy of aggregator and sum-of-customer 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.

Tables ES.1 and ES.2 summarize *accuracy* results for event-type days, for unadjusted and adjusted baselines respectively, showing results for each aggregator and in total. Tables ES.3 and ES.4 summarize *bias* results for unadjusted and adjusted baselines.[1] Expanded tables in the body of the report contain results by industry type and for event days. Major findings include the following:

1. Regarding the accuracy of the *aggregator* method of calculating baselines compared to the *sum-of-customer* method, the results suggest that the aggregator method is more accurate, but not by a wide margin (*e.g.*, compare the two sets of columns in Tables ES.1 and ES.2).
2. 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 (*e.g.*, compare the first columns in Tables ES.3 and ES.4).
3. 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 (*e.g.*, see the third columns in Tables ES.2 and ES.4).
4. Examination of the variability of percent errors of 10-in-10 baselines for *individual customers* illustrates the likely source of greater baseline errors in sum-of-customer baselines compared to aggregator baselines. For example, the *range* of errors is greater for industrial customers than for commercial customers, with a number of large over-stated baselines, although the greatest errors were found to be generally associated with the smallest customers.
5. The performance of the alternative baseline methods on *event days*, in terms of accuracy and bias, appears qualitatively similar to their performance on the *event-type* days summarized above.
6. Analysis in this study revealed no evidence of systematic increases in pre-event consumption on event days that would be indicative of attempts to *game* the adjusted baseline. Only one case was found, for one industrial customer of one aggregator, in which hourly usage rose unusually in the four hours prior to one event, possibly indicating an attempt to increase the baseline from which the load impact would be measured.

[1] Note that by the definition of baseline error used in this study, *positive* errors represent downward biases (*i.e.*, the baseline being tested under-states the true baseline), while *negative* errors represent upward biases (*i.e.*, the baseline being tested over-states the true baseline).

Table ES.1 Accuracy of *Unadjusted* Baselines

		Aggregator			*Sum of Customers*		
		Unadjusted			Unadjusted		
Agg.	Industry	3-in-10	5-in-10	10-in-10	3-in-10	5-in-10	10-in-10
1	Total	0.052	0.064	0.087	0.055	0.056	0.086
2	Total	0.069	0.078	0.106	0.065	0.072	0.106
3	Total	0.046	0.053	0.075	0.068	0.050	0.075
4	Total	0.040	0.036	0.036	0.108	0.080	0.036
All	TOTAL	0.049	0.055	0.076	0.076	0.060	0.076

Table ES.2 Accuracy of *Adjusted* Baselines

		Aggregator					*Sum of Customers*				
		Symmetric Adjustment			*Upward-only*		Symmetric Adjustment			*Upward-only*	
Agg.	Level	3-in-10	5-in-10	10-in-10	5-in-10	10-in-10	3-in-10	5-in-10	10-in-10	5-in-10	10-in-10
1	Total	0.023	0.024	0.023	0.023	0.023	0.038	0.029	0.023	0.048	0.024
2	Total	0.027	0.030	0.029	0.042	0.034	0.042	0.038	0.029	0.053	0.035
3	Total	0.021	0.020	0.020	0.025	0.020	0.044	0.035	0.031	0.072	0.032
4	Total	0.032	0.033	0.031	0.040	0.030	0.073	0.058	0.030	0.108	0.053
All	TOTAL	0.025	0.026	0.025	0.031	0.026	0.050	0.041	0.031	0.076	0.037

Table ES.3 Bias of *Unadjusted* Baselines

		Aggregator			*Sum of Customers*		
		Unadjusted			Unadjusted		
Agg.	Industry	3-in-10	5-in-10	10-in-10	3-in-10	5-in-10	10-in-10
1	Total	4.02%	5.10%	7.95%	-1.26%	1.62%	7.83%
2	Total	3.19%	4.83%	9.31%	-0.28%	2.23%	9.31%
3	Total	4.22%	5.39%	9.30%	0.59%	2.77%	9.21%
4	Total	0.89%	1.72%	4.97%	-2.78%	-0.50%	4.62%
All	TOTAL	3.11%	4.50%	7.72%	-1.01%	1.35%	7.67%

Table ES.4 Bias of *Adjusted* Baselines

		Aggregator					*Sum of Customers*				
		Symmetric Adjustment			*Upward-only Adjustment*		Symmetric Adjustment			*Upward-only Adjustment*	
Agg.	Level	3-in-10	5-in-10	10-in-10	5-in-10	10-in-10	3-in-10	5-in-10	10-in-10	5-in-10	10-in-10
1	Total	-0.37%	-0.11%	0.45%	-0.11%	0.45%	-2.62%	-1.43%	0.64%	-3.46%	-0.41%
2	Total	-0.02%	-0.24%	0.17%	-0.80%	-0.32%	-3.16%	-2.27%	0.43%	-4.01%	-0.31%
3	Total	-0.41%	-0.58%	0.06%	-0.67%	-0.02%	-1.98%	-1.35%	1.38%	-2.68%	0.59%
4	Total	-0.46%	-0.23%	-0.29%	-1.56%	-0.80%	-2.32%	-1.72%	-0.42%	-4.65%	-1.74%
All	TOTAL	-0.38%	-0.29%	0.06%	-1.00%	-0.32%	-2.34%	-1.56%	0.64%	-3.77%	-0.43%

1. Introduction

This report summarizes the results of a baseline analysis that was undertaken as part of the statewide ex-post evaluation of the Aggregator demand response programs. The objective of the baseline analysis was to assess the relative accuracy and bias of several alternative methods for calculating baselines for measuring load impacts for settlement.

The original scope of work involved analysis to address three baseline issues regarding PG&E's Aggregator Managed Portfolio ("AMP") program:

1. Compare the accuracy in measuring load reductions of two alternative methods—Test whether estimating load impacts by comparing actual *aggregate* program loads during an event to an *aggregator* baseline, or to a baseline constructed as the sum of individual *customer-specific* baselines, is more accurate in measuring load reductions.
2. Evaluate whether *morning adjustments* to the 3-in-10 baseline actually improve the bias of the baseline (*i.e.*, the tendency of the calculated baseline to understate or overstate the "true" baseline).
3. Test whether *gaming* is successfully avoided.

The scope was expanded to consider a number of additional baselines and adjustment mechanisms, including the following:

4. Evaluate and compare the accuracy of the following baselines using *day-of adjustment*:
    a. aggregated 3-in-10,
    b. individual 3-in-10,
    c. aggregated 5-in-10,
    d. individual 5-in-10,
    e. aggregated 10-in-10, and
    f. individual 10-in-10.

    The adjustment to be used should be the one AMP currently uses in 2008; that is, the ratio of a) the average load of the 4 hours preceding the event to b) the average load of the same 4 hours of the baseline days.

5. Evaluate the effects of *upward-only* day-of adjustment vs. symmetric day-of adjustment on a baseline. The baseline models to be studied include:
    a. 5-in-10 with a *symmetric* adjustment vs. 5-in-10 with an *upward-only* adjustment, and
    b. 10-in-10 with a *symmetric* adjustment vs. 10-in-10 with an *upward-only* adjustment.
6. Evaluate the effects of allowing the option of symmetric adjustment on:
    a. 10-in-10 (*i.e.*, 10-in-10 unadjusted vs 10-in-10 with a symmetric adjustment), and
    b. 5-in-10 (*i.e.*, 5-in-10 unadjusted vs 5-in-10 with a symmetric adjustment).

2. Data

2.1 Customers

We used data for nearly all of the customers that were nominated by each of four of the five AMP aggregators (one aggregator had only one customer) for the relevant months during 2008. Given the interest in adjusted baselines and gaming for those customers who selected the adjusted baseline option in 2008, some portions of the analysis were conducted separately by customers' choice of adjusted baseline. In addition, to examine potential differences in baseline performance between weather-sensitive and non-weather sensitive customers, we constructed sub-groups of customer types based on their categorization within the standard eight industry groups used in load impact evaluations. The customer types are combinations of the eight standard industry types, and are designed to differentiate between "Industrial-type" customers that are likely to be relatively non-weather sensitive (Industry types 1-3, which include manufacturing, construction, wholesale trade and other utilities), and "Commercial-type" customers, which are likely to be relatively more weather sensitive (Industry types 4, 5, and 7, which include retail stores, offices, services, etc.). Schools (6) were treated separately due to their unique scheduling differences during the summer period.

The number of customers included in the analysis, and their industry type and usage characteristics are shown in Table 1. Aggregators 1 and 2 have relatively large shares of commercial customers, while aggregators 3 and 4 have large shares of industrial customers, and some schools. Aggregators 2 and 4 had a substantial share of customers accept the adjusted baseline option.

Table 1. Characteristics of AMP Baseline Customers

		Count			Max kW			Ind. Type
Agg	Ind. Group	Adj. BL	No Adj	Total	Adj. BL	No Adj	Total	% of Total
1	1		84	84		73,762	73,762	55%
	2		70	70		59,562	59,562	45%
	Total		154	154		133,323	133,323	
2								
	1	18	17	35	16,164	19,088	35,252	29%
	2	81	2	83	87,080	1,136	88,217	71%
	Total	99	19	118	103,244	20,224	123,469	
3								
	1	10	118	128	9,226	212,553	221,778	83%
	2	2	28	30	3,916	31,192	35,108	13%
	3	3	11	14	581	9,556	10,138	4%
	Total	15	157	172	13,723	253,301	267,024	
4								
	1	28	40	68	35,316	91,191	126,507	73%
	2	29	4	33	15,348	9,562	24,911	14%
	3	3		3	22,670		22,670	13%
	Total	60	44	104	73,334	100,753	174,087	
ALL								
	1	56	259	315	60,706	396,593	457,298	66%
	2	112	104	216	106,345	101,452	207,797	30%
	3	6	11	17	23,251	9,556	32,807	5%
	Total	174	374	548	190,301	507,601	697,903	

2.2 Events

Given the relatively small number of AMP events (one actual event for one of the aggregators, and four test events for a mix of aggregators), and the availability of a number of days of relatively high PG&E system load and temperatures, we conducted much of the baseline analysis for ten *event-type* days during the May to September period. These are shown in Table 2 (actual and test events are shown in highlight). The simulated events were assumed to be five hours in length, from hours ending 14 through 18. Morning adjustments were made using consumption in the four hours prior to the "event." In addition to the event-type days, we also examined baseline performance on the two actual events (events 1 and 3) for which all aggregators were called. The analysis of gaming was also conducted using data for the two event days.

Table 2. Event-type Days

	AMP Ld (HE 14)	wCDD	DOW	Max temp	Event-like day
12-May-08	479,979	1.5	1		
13-May-08	498,251	4.2	2		
14-May-08	534,785	9.8	3		
15-May-08	554,240	17.6	4	99.6	
16-May-08	495,091	18.8	5	101.4	
19-May-08	516,584	9.8	1		
20-May-08	512,840	6.6	2		
21-May-08	490,420	1.3	3		
22-May-08	483,425	1.6	4		
23-May-08	453,046	1.0	5		
26-May-08	327,109	-	1		
27-May-08	464,516	0.2	2		
28-May-08	480,879	0.1	3		
29-May-08	481,636	0.7	4		
30-May-08	470,445	1.5	5		
2-Jun-08	486,074	2.8	1		
3-Jun-08	501,945	3.2	2		
4-Jun-08	497,220	1.9	3		
5-Jun-08	514,825	4.2	4		
6-Jun-08	479,156	3.3	5		
9-Jun-08	529,353	11.8	1	97	
10-Jun-08	519,318	9.6	2		
11-Jun-08	508,973	8.7	3		
12-Jun-08	534,801	11.0	4		
13-Jun-08	515,111	11.6	5	96	
16-Jun-08	506,289	8.0	1		
17-Jun-08	528,354	9.2	2	95	
18-Jun-08	534,248	12.1	3		
19-Jun-08	550,430	14.5	4	98	
20-Jun-08	**544,268**	**19.5**	**5**	**103.2**	**X**
23-Jun-08	511,975	8.1	1		
24-Jun-08	530,964	8.3	2		
25-Jun-08	522,482	7.1	3		
26-Jun-08	528,335	7.4	4		
27-Jun-08	503,760	10.8	5		
30-Jun-08	505,490	8.0	1		
1-Jul-08	510,351	7.7	2		
2-Jul-08	518,861	8.5	3		
3-Jul-08	514,146	9.7	4		
4-Jul-08	376,261	7.4	5		
7-Jul-08	**566,085**	**17.2**	**1**	**104.2**	**X**
8-Jul-08	**583,342**	**20.9**	**2**	**107.2**	**X**
9-Jul-08	572,732	21.0	3	105.6	
10-Jul-08	**574,970**	**19.2**	**4**	**102.2**	**X**
11-Jul-08	542,147	12.2	5		

	AMP Ld (HE 14)	wCDD	DOW	Max temp	Event-like day
14-Jul-08	548,959	12.0	1		
15-Jul-08	541,164	9.6	2		
16-Jul-08	544,699	9.9	3		
17-Jul-08	540,822	10.2	4		
18-Jul-08	516,688	9.5	5		
21-Jul-08	508,257	5.1	1		
22-Jul-08	531,218	7.3	2		
23-Jul-08	542,347	11.0	3		
24-Jul-08	541,589	10.6	4		
25-Jul-08	520,680	11.4	5		
28-Jul-08	514,269	7.7	1		
29-Jul-08	522,534	8.4	2		
30-Jul-08	526,255	9.3	3		
31-Jul-08	529,726	10.0	4		
1-Aug-08	509,300	9.6	5		
4-Aug-08	518,059	9.4	1		
5-Aug-08	522,927	9.3	2		
6-Aug-08	522,683	10.7	3		
7-Aug-08	521,209	10.4	4		
8-Aug-08	504,878	8.6	5		
11-Aug-08	555,597	13.1	1	97.2	
12-Aug-08	548,655	12.9	2	97.8	
13-Aug-08	**572,702**	**15.7**	**3**	**101.6**	**X**
14-Aug-08	555,178	15.4	4	101.2	
15-Aug-08	**554,229**	**16.2**	**5**	**101.8**	**X**
18-Aug-08	547,062	7.2	1		
19-Aug-08	546,683	6.1	2		
20-Aug-08	567,727	8.1	3		
21-Aug-08	574,845	10.6	4		
22-Aug-08	558,487	11.5	5		
25-Aug-08	566,229	12.7	1		
26-Aug-08	574,105	11.7	2		
27-Aug-08	**585,571**	**16.1**	**3**	**101**	**X**
28-Aug-08	**596,961**	**19.3**	**4**	**104.6**	**X**
29-Aug-08	**568,149**	**19.3**	**5**	**103.8**	**X**
1-Sep-08	388,091	7.0	1		
2-Sep-08	553,231	10.8	2		
3-Sep-08	569,000	13.5	3	98.2	
4-Sep-08	**590,351**	**15.8**	**4**	**101.2**	**X**
5-Sep-08	578,943	16.8	5	102.4	

3. Approach

Two general alternative methods for constructing the baseline load for aggregators were examined, in both unadjusted and adjusted forms. These are the following:

1. *Aggregator-level baseline* – In this method, the hourly loads for all of an aggregator's nominated customers are summed, and the resulting aggregator loads are used to identify the highest 3-in-10 (as well as 5-in-10 and 10-in-10) days for each event-type day, and the average loads over the selected days are calculated. The resulting aggregator baselines are then compared to the *actual* aggregator load for each of the event-type days. This is the current baseline approach used for AMP, with the 3-in-10 averaging method.[2]
2. *Sum-of-customer baseline* – In this method, the hourly loads for each of an aggregator's customers are used separately to identify their highest 3-in-10 (or 5-in-10 and 10-in-10) days for each event-type day, the average loads over those three days are calculated, and then the individual customer baseline loads are summed up to produce a (different) aggregator baseline load for each event-type day. The resulting sum-of-customer baselines are then compared to the *actual* aggregator load for each of the event-type days.[3]

Two different methods were used for developing the "true" baselines to which the alternative baseline methods were compared, depending on whether the events being analyzed were *actual* event days or *event-type* days.[4] An advantage of using event-type days that were not actual event days is that consumers' *actual* loads on those days may be used as the *true* baseline for purposes of comparing alternative baselines which are estimated as averages of previous days' loads. In the case of actual events, the true baselines must be estimated, typically using information from regression analyses of customers' loads. For the actual events in this study, we constructed the "true" baseline for each customer as the sum of their observed load and our estimated load impact coefficients from the individual customer regressions described in Volume 1 of the report.[5] The true baseline for each aggregator and sub-group of customers, for each event, was then calculated as the sum of the individual baselines for the relevant customers.

3.1 Baseline performance statistics

For each of the baseline methods, two statistics are calculated to compare the performance of estimated baselines to the true baselines (*e.g.*, the actual load on the event-type day). One statistic measures *accuracy*, while the other measures *bias*, or the tendency of a particular baseline method to under-state or over-state the true baseline.

---

[2] Three of the aggregators offered their customers a choice of an adjusted 3-in-10 baseline for 2008. Otherwise, the program baseline was an unadjusted aggregator-level 3-in-10 baseline.

[3] The primary difference between the two baselines is analogous to the difference between coincident and non-coincident demands. The sum-of-customers baseline adds together each customer's (non-coincident) average of highest three loads in the past ten days, while the aggregator baseline averages each customer's loads over the three (coincident) days that represent the *aggregator's* highest load. It is generally acknowledged that summing each individual customers' highest three loads will tend to produce a higher baseline than if the baseline is based on the highest (diversified) load of the aggregator.

[4] Days on which events were called for only some aggregators were included as event-type days for the aggregators who were not called.

[5] This method is analogous to the approach used to construct program reference loads in the ex post and ex ante load impact evaluations from the observed loads on event days and the estimated program load impacts.

3.1.1 Accuracy

Accuracy is measured using the *relative root mean square error* statistic (RRMSE, sometimes referred to as the Theil U-statistic). The formula for this statistic is the following:

> *U-statistic* = $[(1/n) \sum (e_h)^2]^{1/2} / [(1/n) \sum (L^A{}_h)^2]^{1/2}$,
> where
>
> e_h = $(L^A{}_h - L^P{}_h)$,
> $L^A{}_h$ is *actual* load,
> $L^P{}_h$ is *predicted* (baseline) load,
> *n* is the total number of event days and hours, and
> the sum is across event days and hours, for each aggregator, or sub-group by industry type.

This statistic measures the degree of difference, or error, between the two data series, $L^P{}_h$ and $L^A{}_h$. It 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 0.05 indicates an average 5 percent error in the baseline relative to its mean value.

3.1.2 Bias

The other statistic, which is used to measure the typical *direction* of error, is the *median % error*:

> *Median percentage error* = Median of $(e_h / L^A{}_h)$, across event days and hours, for each aggregator, or sub-group by industry type.

This statistic has been used to measure the *bias* in the baseline load, indicating 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, examining the *distribution* of percent errors provides insight into the full range of baseline errors. Finally, it is important to note that the convention of defining errors, e_h, as the difference between actual and estimated baseline values $(L^A{}_h - L^P{}_h)$, implies that *positive* errors represent downward bias, or under-stated baselines, while *negative* errors represent upward bias, or over-stated baselines.

3.2 Adjusted baselines

Two sets of *adjusted* versions of each of the baseline methods have also been assessed—a *symmetric* adjustment, and an *upward-only* adjustment. In both cases, the adjustments take the form of the ratio of the average load on the event day in the four hours prior to the event, to the average load in the same four hours of the unadjusted baseline, based on the highest three, five, or ten day approaches. The adjustment involves multiplying the unadjusted baseline times the adjustment ratio. The objective of the adjustment is to take advantage of information on customers' usage in the pre-event hours of an event day to improve the accuracy of an unadjusted baseline, which otherwise represents customers' usage on days that may be less extreme in terms of weather conditions than the event day. The symmetric adjustments were limited to no more than a 20 percent increase or decrease from the unadjusted baseline. For the upward-only adjustment, only positive adjustments were made.

4. Results for *Event-type Days*

4.1 Unadjusted baselines

We begin by establishing a reference point of performance results for the alternative *unadjusted* baselines.

4.1.1 Accuracy

Table 3 shows *accuracy* results for unadjusted versions of the three different methods based on the number of days selected for inclusion in the baseline calculation (*e.g.*, 3, 5, or 10), and for the two different methods for calculating aggregate baselines—aggregator and sum-of-customers. Figures 1 and 2 plot the values in Table 4, providing a helpful visual characterization of the patterns of values. The following observations characterize some of the important results:

- For the unadjusted 3-in-10 *aggregator* baseline, shown in the first column, and focusing first on the rows labeled TOTAL for each aggregator, relative errors range from about 4 to 7 percent across the aggregators, with a relative error of 5 percent across all customers.
- For the comparable unadjusted *sum-of-customer* baseline, shown in the first column of the second group of columns, the relative errors are generally similar or somewhat larger for the first three aggregators, and substantially larger for the fourth, with an overall relative error of 7.6 percent.
- Moving across to the 5-in-10 and 10-in-10 columns, the relative errors for the aggregator baseline generally *increase* with the number of days included in the baseline average (the exception is the fourth aggregator, where the errors remain relatively constant).
- Comparing results by industry type, the findings suggest that the relative errors for *commercial* customers generally increase with the number of days included in the baseline average, but that the patterns of relative errors for *industrial* customers differ by aggregator and level of aggregation (aggregator or sum of customers).
- For three of the aggregators, the relative errors of the *aggregator-level* baselines for commercial customers are greater than for industrials; however, for the *sum-of-customer* baselines, the industrial group generally has greater relative errors.
- For both methods, *schools* have among the highest relative errors.
- Relative errors for an aggregation across *all* customers, shown in the last set of rows, are generally consistent with the aggregator-level results; relative errors are somewhat smaller for the aggregator baseline than for the sum-of-customer baseline, except for the 10-in-10, where the relative errors are the same.

Table 3. Accuracy of Unadjusted Baselines
(Relative root mean square error, or Theil U-statistic)

		Aggregator			*Sum of Customers*		
		Unadjusted			**Unadjusted**		
Aggregator	**Industry**	**3-in-10**	**5-in-10**	**10-in-10**	**3-in-10**	**5-in-10**	**10-in-10**
1	Industry	0.039	0.045	0.058	0.065	0.044	0.057
	Commercial	0.058	0.073	0.101	0.047	0.061	0.100
	TOTAL	0.052	0.064	0.087	0.055	0.056	0.086
2	Industry	0.126	0.122	0.137	0.168	0.144	0.134
	Commercial	0.062	0.073	0.103	0.046	0.062	0.103
	TOTAL	0.069	0.078	0.106	0.065	0.072	0.106
3	Industry	0.045	0.051	0.074	0.068	0.049	0.074
	Commercial	0.066	0.073	0.096	0.045	0.057	0.095
	Schools	0.099	0.110	0.141	0.085	0.101	0.141
	TOTAL	0.046	0.053	0.075	0.068	0.050	0.075
4	Industry	0.038	0.032	0.028	0.112	0.082	0.028
	Commercial	0.041	0.049	0.069	0.027	0.038	0.069
	Schools	0.080	0.079	0.105	0.081	0.077	0.101
	TOTAL	0.040	0.036	0.036	0.108	0.080	0.036
All	Industry	0.046	0.050	0.068	0.082	0.060	0.068
	Commercial	0.060	0.072	0.100	0.046	0.060	0.100
	Schools	0.085	0.087	0.114	0.082	0.083	0.111
	TOTAL	0.049	0.055	0.076	0.076	0.060	0.076

Figure 1. Accuracy of *Unadjusted* Baselines – *Aggregator*

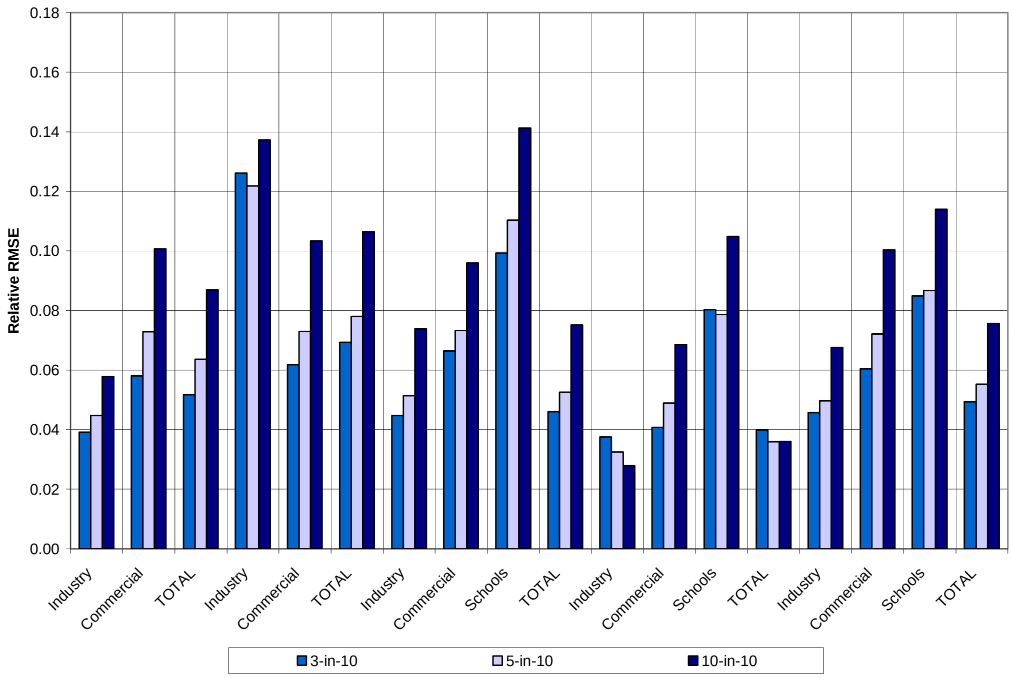
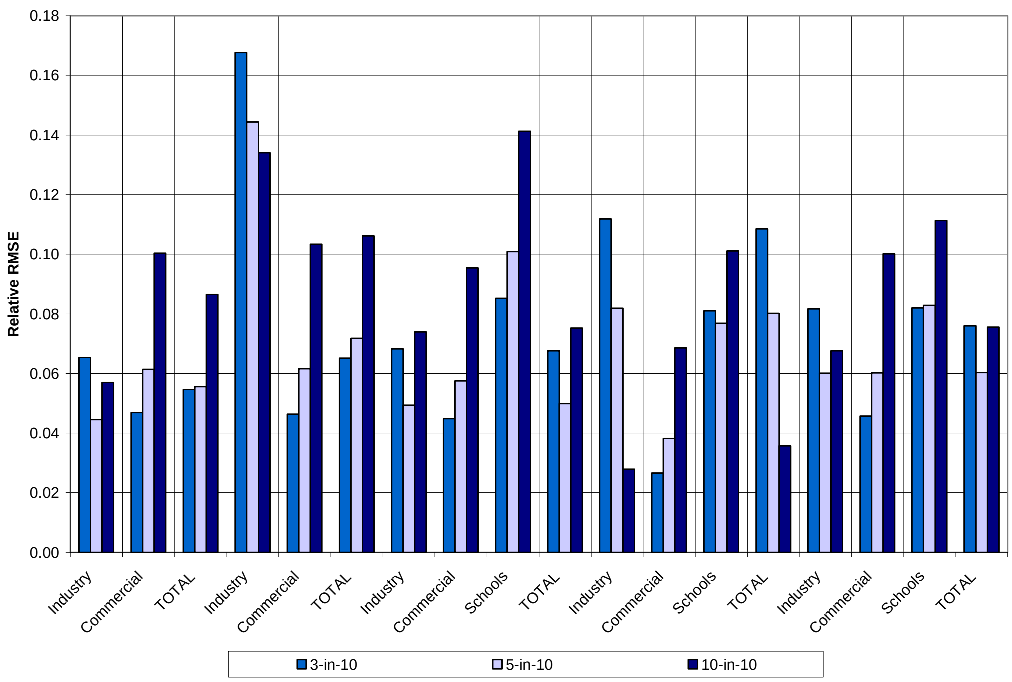


Figure 2. Accuracy of *Unadjusted Baselines – Sum-of-Customers*



4.1.2 Bias

Table 4, and Figures 3 and 4 present comparable results for bias, showing the median % errors across event-type days and hours, by aggregator and across all customers. 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 in the TOTAL rows in the first column are positive, indicating that the unadjusted 3-in-10 aggregator baseline is typically biased downward (*i.e.*, typically *under-states* the true baseline) for three of the four aggregators by about 4 percent, and for the fourth aggregator by less than 1 percent.
- In contrast, the sum-of-customer method produces quite small biases for two of the aggregators, and somewhat *over-stated* baselines for the other two.
- Looking across methods, 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 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* (typically ranging from 5 to 10 percent across number of days in the baseline) than for *industrial* customers.
- For most aggregators, the unadjusted *sum-of-customer* baseline for industrial customers tends to *over-state* the true baseline, particularly for the 3-in-10 and 5-in-10 methods, though those results are reversed for the 10-in-10 method (*i.e.*, baselines are over-stated).

Table 4. Bias of *Unadjusted* Baselines
(Median percent errors)

		Aggregator			*Sum of Customers*		
		Unadjusted			Unadjusted		
Aggregator	**Industry**	**3-in-10**	**5-in-10**	**10-in-10**	**3-in-10**	**5-in-10**	**10-in-10**
	Industry	2.93%	3.20%	5.06%	-4.90%	-2.06%	4.94%
	Commercial	5.42%	6.73%	8.93%	4.30%	5.67%	8.93%
1	TOTAL	4.02%	5.10%	7.95%	-1.26%	1.62%	7.83%
	Industry	-1.32%	0.38%	6.19%	-9.41%	-5.53%	5.41%
	Commercial	5.35%	7.23%	10.29%	3.57%	5.68%	10.29%
2	TOTAL	3.19%	4.83%	9.31%	-0.28%	2.23%	9.31%
	Industry	2.98%	3.96%	7.76%	-2.92%	0.07%	7.75%
	Commercial	5.04%	5.08%	9.03%	1.35%	3.03%	9.03%
	Schools	6.47%	9.26%	13.21%	5.47%	8.72%	13.21%
3	TOTAL	4.22%	5.39%	9.30%	0.59%	2.77%	9.21%
	Industry	-2.67%	-2.14%	0.38%	-10.69%	-7.52%	0.33%
	Commercial	4.56%	5.36%	7.37%	2.47%	4.25%	7.37%
	Schools	0.88%	3.32%	9.85%	-1.21%	1.27%	9.33%
4	TOTAL	0.89%	1.72%	4.97%	-2.78%	-0.50%	4.62%
	Industry	-0.24%	0.40%	4.01%	-7.49%	-5.04%	3.89%
	Commercial	4.84%	5.93%	8.58%	2.75%	4.52%	8.50%
	Schools	4.65%	6.79%	11.99%	3.04%	5.18%	11.66%
All	TOTAL	3.11%	4.50%	7.72%	-1.01%	1.35%	7.67%

Figure 3. Bias of *Unadjusted* Baselines – *Aggregator*

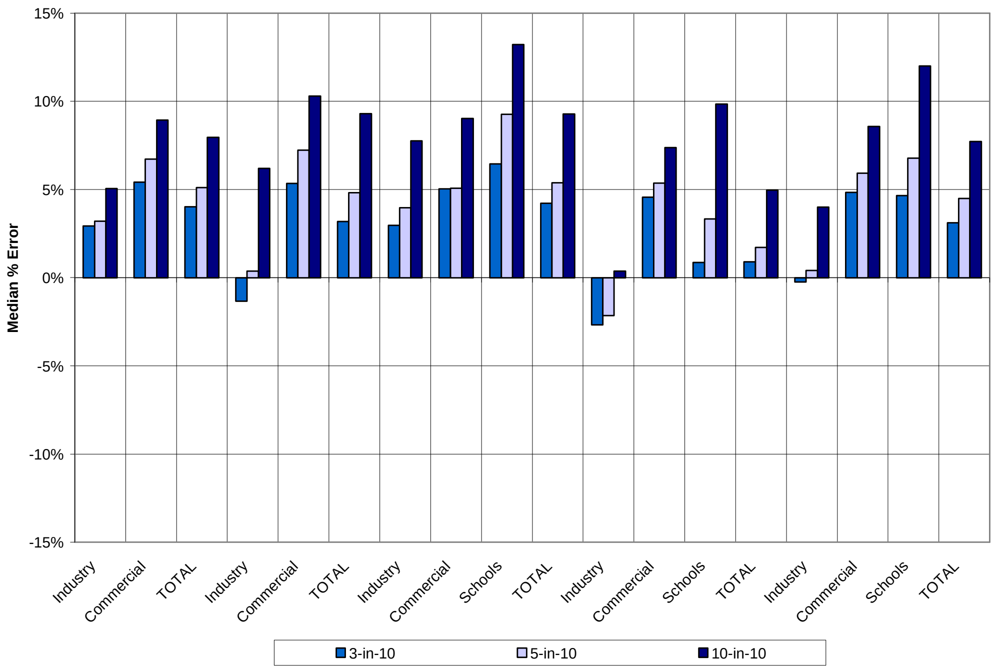
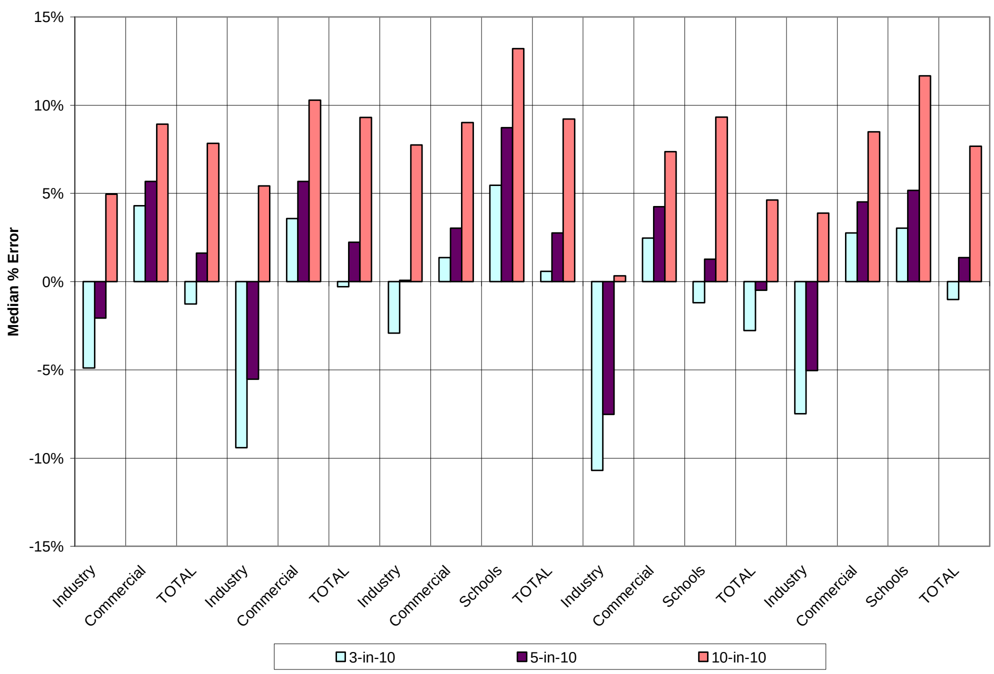


Figure 4. Bias of *Unadjusted* Baselines – *Sum-of-Customers*



4.2 Adjusted baselines

4.2.1 Accuracy

Table 5, and Figures 5 and 6 report accuracy results for the various alternative adjustment methods, for the aggregator and sum-of-customer baselines. Key findings include the following:

- Focusing first on the TOTAL rows, the symmetric morning adjustment generally improves baseline accuracy substantially, reducing relative errors by half or more in many cases compared to the unadjusted baselines.
- For the *aggregator* baseline in particular, the relative errors of the adjusted baselines are very similar across the number of days included in the baseline, even for the upward-only adjustment method.
- For the sum-of-customer baseline, differences in relative accuracy are greater, with the adjusted 10-in-10 baseline generally showing the greatest accuracy, and the upward-only adjustment alternative generally producing somewhat larger relative errors than the corresponding symmetric adjustment.
- 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, and the adjusted baselines for *schools* are the least accurate.
- The adjusted *aggregator* baselines are generally more accurate than the *sum-of-customer* baselines, especially so for industrial customers and for the upward-only adjustments, where the relative errors of the sum-of-customer baselines are often substantially larger than for the aggregator baselines.
- The two upward-only adjustments reduce the accuracy of the aggregator baseline only slightly compared to the symmetric adjustments, but reduce the accuracy of the sum-of-customers baseline more substantially for some aggregators and industry types.
- Results across all customers confirm those at the aggregator level.

Table 5. Accuracy of *Adjusted* Baselines
(Relative root mean square error, or Theil U-statistic)

		Aggregator					*Sum of Customers*				
		Symmetric Adjustment			*Upward-only*		Symmetric Adjustment			*Upward-only*	
Agg.	**Industry**	**3-in-10**	**5-in-10**	**10-in-10**	**5-in-10**	**10-in-10**	**3-in-10**	**5-in-10**	**10-in-10**	**5-in-10**	**10-in-10**
1	Industry	0.028	0.027	0.023	0.026	0.023	0.054	0.040	0.031	0.072	0.034
	Commercial	0.020	0.022	0.022	0.021	0.022	0.022	0.019	0.016	0.024	0.015
	TOTAL	0.023	0.024	0.023	0.023	0.023	0.038	0.029	0.023	0.048	0.024
2	Industry	0.060	0.055	0.050	0.118	0.084	0.120	0.102	0.069	0.163	0.100
	Commercial	0.022	0.027	0.026	0.027	0.026	0.026	0.026	0.022	0.027	0.021
	TOTAL	0.027	0.030	0.029	0.042	0.034	0.042	0.038	0.029	0.053	0.035
3	Industry	0.021	0.020	0.020	0.025	0.020	0.044	0.036	0.032	0.073	0.032
	Commercial	0.020	0.018	0.018	0.018	0.018	0.024	0.021	0.018	0.026	0.018
	Schools	0.046	0.044	0.037	0.051	0.037	0.044	0.036	0.038	0.051	0.040
	TOTAL	0.021	0.020	0.020	0.025	0.020	0.044	0.035	0.031	0.072	0.032
4	Industry	0.031	0.032	0.029	0.040	0.031	0.076	0.060	0.030	0.111	0.054
	Commercial	0.014	0.013	0.013	0.013	0.013	0.013	0.014	0.012	0.020	0.011
	Schools	0.063	0.065	0.071	0.056	0.037	0.051	0.046	0.036	0.076	0.058
	TOTAL	0.032	0.033	0.031	0.040	0.030	0.073	0.058	0.030	0.108	0.053
All	Industry	0.026	0.025	0.025	0.032	0.026	0.055	0.044	0.033	0.084	0.040
	Commercial	0.021	0.024	0.023	0.024	0.023	0.024	0.023	0.020	0.026	0.019
	Schools	0.060	0.061	0.065	0.055	0.037	0.049	0.044	0.037	0.071	0.054
	TOTAL	0.025	0.026	0.025	0.031	0.026	0.050	0.041	0.031	0.076	0.037

Figure 5. Accuracy of *Adjusted* Baselines – *Aggregator*

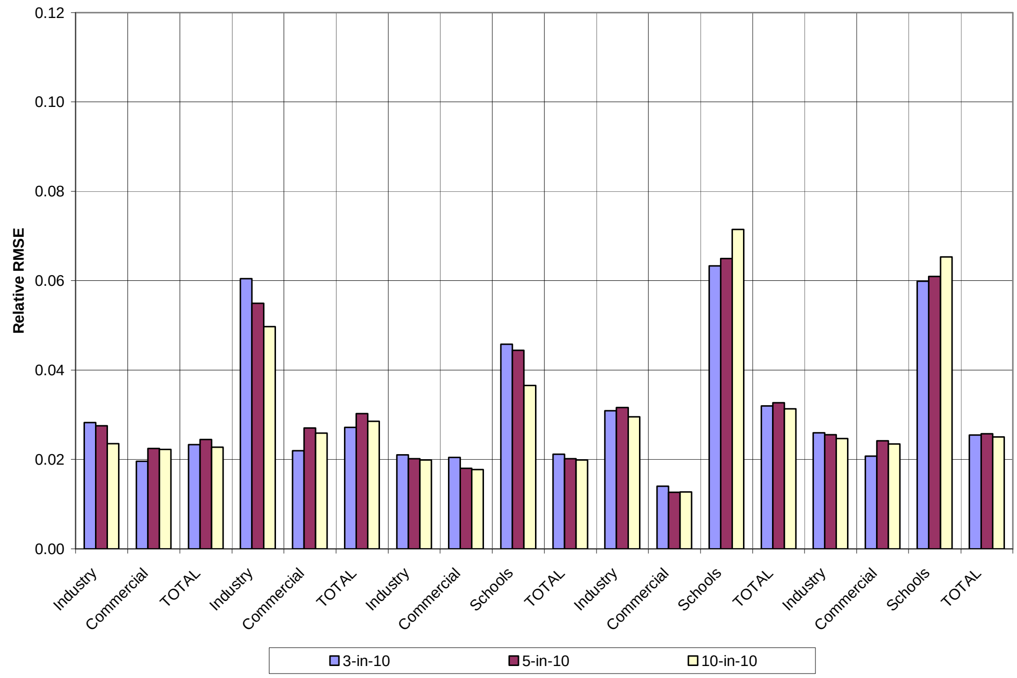
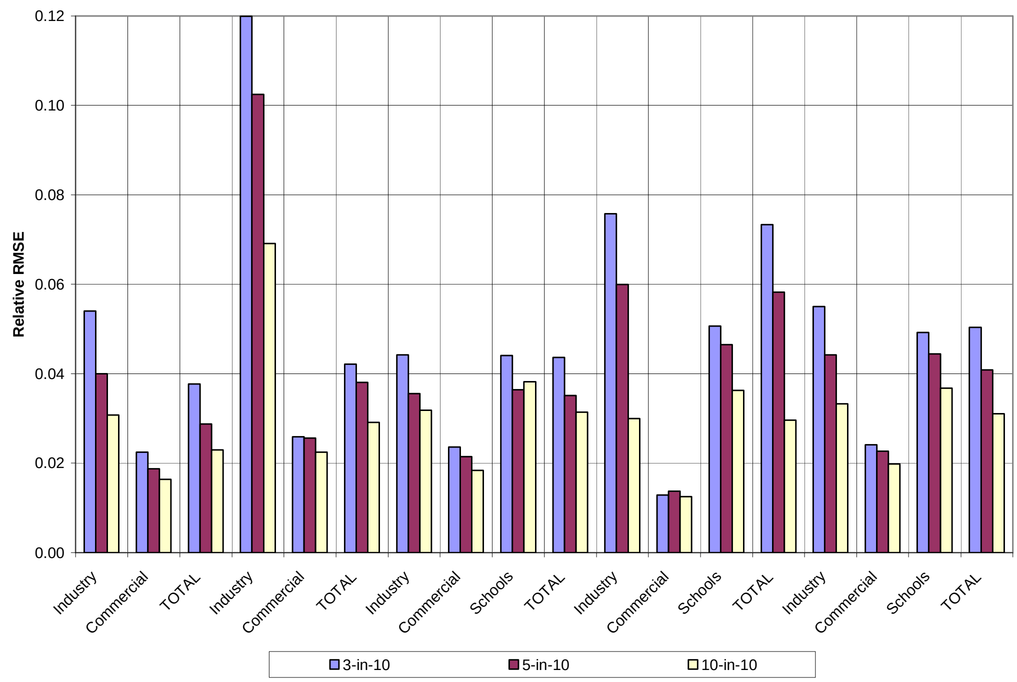


Figure 6. Accuracy of *Adjusted* Baselines – *Sum of Customer*



4.2.2 Bias

Table 6, and Figures 7 and 8 report bias results for the alternative adjustment methods. Key results are the following:

- At the TOTAL level, and looking first at the first column, the morning adjustments generally convert the typical 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 as measured by the median percent error). The morning adjustments for the sum of customer baselines generally increase the upward bias by more than for the aggregator baselines. (Compare Figures 7 and 8 to Figures 3 and 4.)
- Looking across columns as the number of days included in the baseline increases, the extent of upward bias appears to decrease, to the point that in most cases the bias of the adjusted 10-in-10 baseline typically shows a small under-statement. Across all customers, the median % error is near zero. (See the value of 0.06% in the last row of the third column.)
- For the *sum-of-customers* baseline, the median bias across all customers changes somewhat more than for the aggregator baseline, from an upward bias of 2.34% for the adjusted 3-in-10, to a downward bias of 0.64% for the adjusted 10-in-10.
- Looking *across industry types*, there are few consistent patterns for the aggregator baselines, although the end result is that the adjusted 10-in-10 baseline has a somewhat smaller bias.
- For the sum-of-customers method, the adjusted baselines for *industrial* customers are generally biased upward (*i.e.*, the median % errors take on larger negative values) by more than those for commercial customers, but this feature is reduced by moving from the 3-in-10 to 10-in-10 baseline.
- The upward bias and difference between industrial and commercial customer types is particularly evident for the upward-only adjustment for the 5-in-10 baseline (*e.g.*, see the second to last columns in each group, and the second to last bars in each group of bars in Figures 7 and 8).
- The adjusted versions of the 5-in-10 and 10-in-10 have smaller biases (frequently less than 1 percent) than the unadjusted versions, which have median relative errors suggesting typical understated baselines of about 5 percent for the 5-in-10, and 8 to 9 percent for the 10-in-10.
- The upward-only adjustments to the 5-in-10 and 10-in-10 baselines *increase* the bias of the aggregator baseline modestly, particularly for the fourth aggregator, but increase the bias more substantially for the sum-of-customers baseline, not unexpectedly producing greater *upward bias*, which for the 5-in-10 is around 4 percent.

Table 6. Bias of *Adjusted* Baselines
(Median percent errors)

		Aggregator					*Sum of Customers*				
		Symmetric Adjustment			***Upward-only Adjustment***		**Symmetric Adjustment**			***Upward-only Adjustment***	
Agg.	**Industry**	**3-in-10**	**5-in-10**	**10-in-10**	**5-in-10**	**10-in-10**	**3-in-10**	**5-in-10**	**10-in-10**	**5-in-10**	**10-in-10**
	Industry	-0.30%	-0.28%	0.12%	-0.28%	0.12%	-4.04%	-2.17%	1.14%	-5.55%	-0.79%
	Commercial	-0.50%	0.04%	0.55%	0.04%	0.55%	-1.41%	-0.66%	0.60%	-1.83%	0.08%
1	TOTAL	-0.37%	-0.11%	0.45%	-0.11%	0.45%	-2.62%	-1.43%	0.64%	-3.46%	-0.41%
	Industry	-0.06%	-0.24%	1.92%	-2.49%	1.42%	-5.05%	-3.42%	1.31%	-8.45%	-1.17%
	Commercial	-0.01%	-0.24%	-1.16%	-0.24%	-1.16%	-1.43%	-1.16%	0.10%	-1.91%	-0.15%
2	TOTAL	-0.02%	-0.24%	0.17%	-0.80%	-0.32%	-3.16%	-2.27%	0.43%	-4.01%	-0.31%
	Industry	-0.79%	-0.44%	0.00%	-0.74%	0.00%	-3.34%	-1.69%	2.51%	-5.67%	0.59%
	Commercial	-0.32%	-0.38%	-0.03%	-0.38%	-0.03%	-1.30%	-0.79%	0.74%	-1.92%	0.35%
	Schools	-0.62%	-0.79%	0.18%	-1.22%	-0.01%	-1.74%	-0.94%	1.78%	-1.80%	1.08%
3	TOTAL	-0.41%	-0.58%	0.06%	-0.67%	-0.02%	-1.98%	-1.35%	1.38%	-2.68%	0.59%
	Industry	-0.91%	-0.70%	-0.29%	-3.61%	-1.64%	-7.08%	-5.28%	-1.66%	-10.68%	-5.06%
	Commercial	-0.46%	-0.12%	0.16%	-0.12%	0.16%	-0.76%	-0.41%	0.60%	-1.34%	0.01%
	Schools	-0.13%	-0.02%	-0.98%	-3.15%	-1.53%	-1.49%	-0.94%	-1.23%	-4.65%	-2.44%
4	TOTAL	-0.46%	-0.23%	-0.29%	-1.56%	-0.80%	-2.32%	-1.72%	-0.42%	-4.65%	-1.74%
	Industry	-0.44%	-0.35%	0.35%	-2.22%	-0.73%	-4.56%	-2.76%	0.97%	-7.29%	-1.76%
	Commercial	-0.31%	-0.23%	-0.03%	-0.23%	-0.03%	-1.05%	-0.76%	0.53%	-1.79%	0.09%
	Schools	-0.43%	-0.58%	-0.51%	-1.88%	-0.71%	-1.72%	-0.94%	0.80%	-3.00%	-0.58%
All	TOTAL	-0.38%	-0.29%	0.06%	-1.00%	-0.32%	-2.34%	-1.56%	0.64%	-3.77%	-0.43%

Figure 7. Bias of *Adjusted* Baselines – *Aggregator*

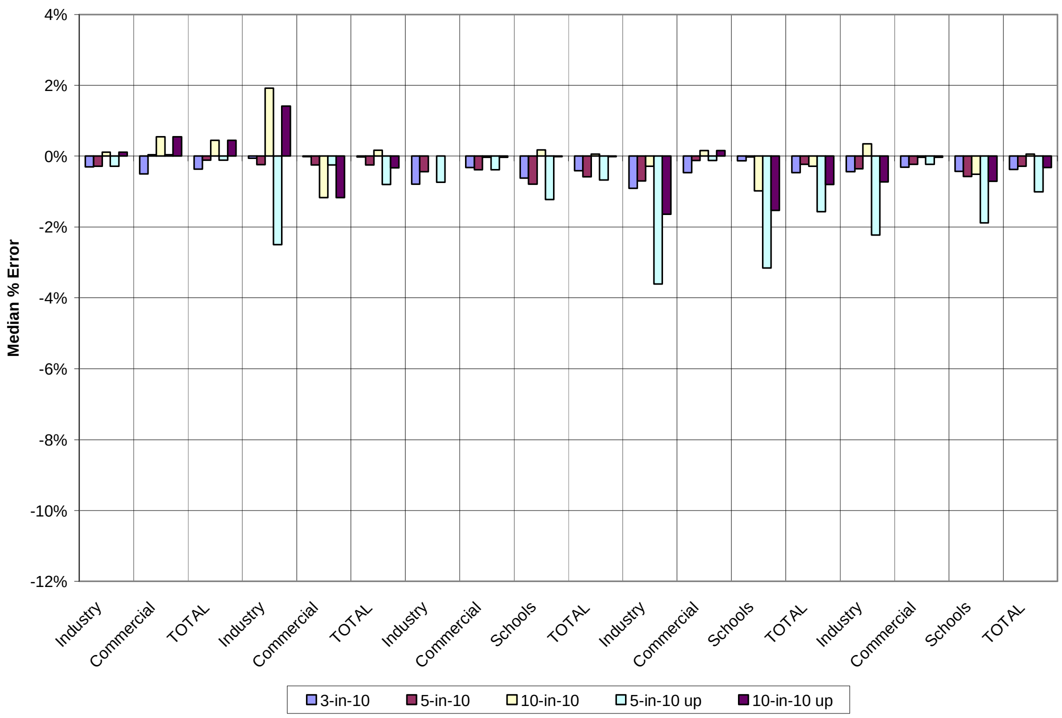
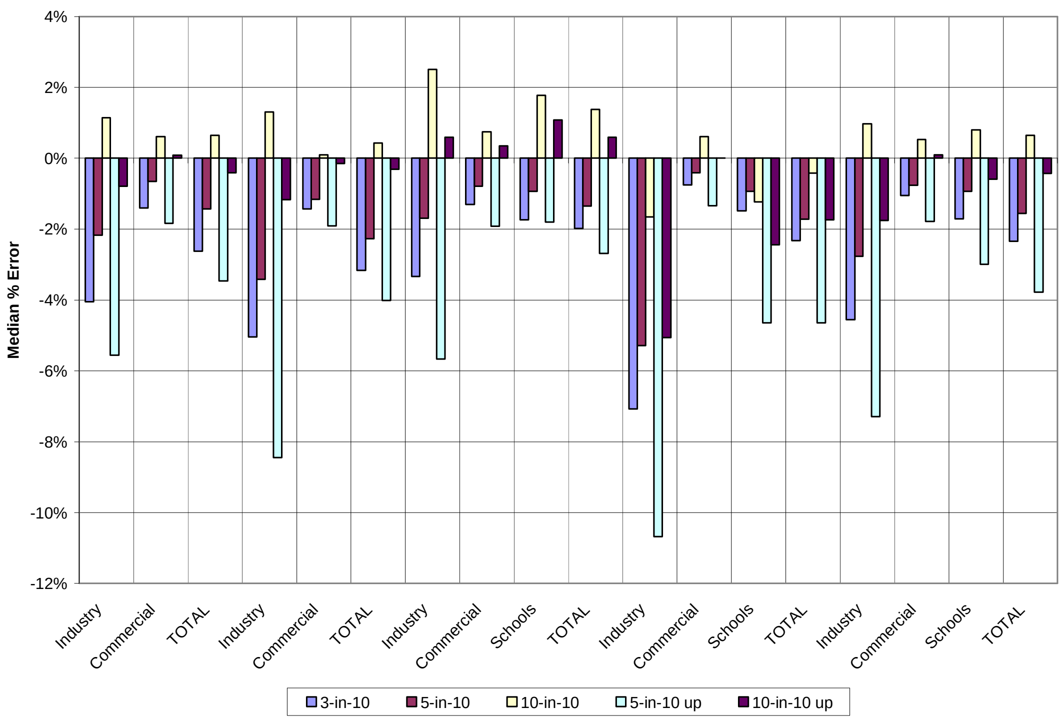


Figure 8. Bias of *Adjusted* Baselines – *Sum of Customer*



4.2.3 Conclusions – Event-type days

The variability of the above results across aggregators and customer types suggests that baseline performance depends on a number of factors, and that conclusions regarding the performance of particular baseline methods are not definitive in all cases.[6] Nevertheless, some reasonably consistent findings may be reported on the key issues of interest to the utilities. These include the following:

1. An aggregator baseline approach appears to generally provide a more accurate estimate of the true baseline than a sum-of-customer baseline that is constructed as the sum of individual customer baselines. For unadjusted baselines, the difference in accuracy is modest, particularly as the number of days included in the baseline increases; for adjusted baselines, the difference is somewhat greater, but declines with the number of days included in the baseline.
2. An unadjusted aggregator baseline approach typically *under-states* the true baseline by about 4 percent (for the 3-in-10 baseline). In contrast, the unadjusted 3-in-10 sum-of-customer baseline has a small upward bias (1 percent). As the number of days included in the baseline increases, both methods produce larger downward biases, converging to a median percent error of nearly 8 percent for the unadjusted 10-in-10 baseline.
3. *Morning adjustments* to the 3-in-10 baseline improve both the accuracy and bias of the unadjusted version, particularly for the aggregator method. Adjusted versions of the sum-of-customer baseline produce a larger upward bias than the aggregator baseline. However, the biases of both methods are smallest and reasonably close together for the adjusted 10-in-10 baseline.
4. The *accuracies* of adjusted versions of the 3-in-10, 5-in-10, and 10-in-10 aggregator baselines are quite similar, and somewhat more accurate than the comparable sum-of-customer baselines. However, the accuracies of the two methods appear to converge somewhat as the number of days included in the baseline increases to 10-in-10.
5. The *biases* of adjusted versions of the 3-in-10, 5-in-10, and 10-in-10 aggregator baselines are also quite similar, and considerably smaller than the comparable sum-of-customer baselines. However, the biases of the two methods also appear to converge somewhat as the number of days included in the baseline increases to 10-in-10.
6. The *upward-only* adjustments to the 5-in-10 and 10-in-10 baselines *increase* the bias of the aggregator baseline modestly, particularly for the fourth aggregator, but increase the bias more substantially for the sum-of-customers baseline, not unexpectedly producing greater *upward bias*.

4.3 Distributions of relative errors

While the median percent error provides a useful indicator of the tendency of a particular baseline method to under-state or over-state the true baseline, the single median value can mask a potentially wide range of relative (percent) errors across event days and hours. This section illustrates several features of the range of baseline errors. The first part of the section focuses on results at the aggregator/industry-type level. The second part shows underlying results at the

[6] Additional calculations made but not reported here suggest that baseline performance can also depend on the nature and timing of events, such as whether they are isolated events that follow several days of non-event days, or are events that occur following one or more events, thus pushing back the days included in the baseline calculation farther away from the event day.

individual customer level. Given the relatively strong performance of the adjusted 10-in-10 baseline report above, the results in this section use errors calculated for that baseline method.

4.3.1 Distributions by aggregator and industry type

The following figures show the relationship between the relative (percentage) errors of the adjusted *aggregator* and *sum of customer* baselines (using the adjusted 10-in-10 baseline), where each point represents the average percent error across event hours for an aggregator, industry type and event day.[7] The values are sorted according to the value for the aggregator baseline. Figure 9 shows values across all industry types. For the most part, those errors range from –5 percent (indicating a five percent over-statement) to +5 percent, with a handful of outliers. The percent errors of the sum-of-customer baseline appear on average to lie above the values for the aggregator baseline (thus indicating a somewhat *higher* baseline), which is consistent with the difference in overall medians (0.64 percent for the sum-of-customers, versus essentially zero for the aggregator).

Figure 9. Average Event-Day % Errors for *Adjusted 10-in-10* Aggregator and Sum-of-Customer Baselines – *All Industry Types*

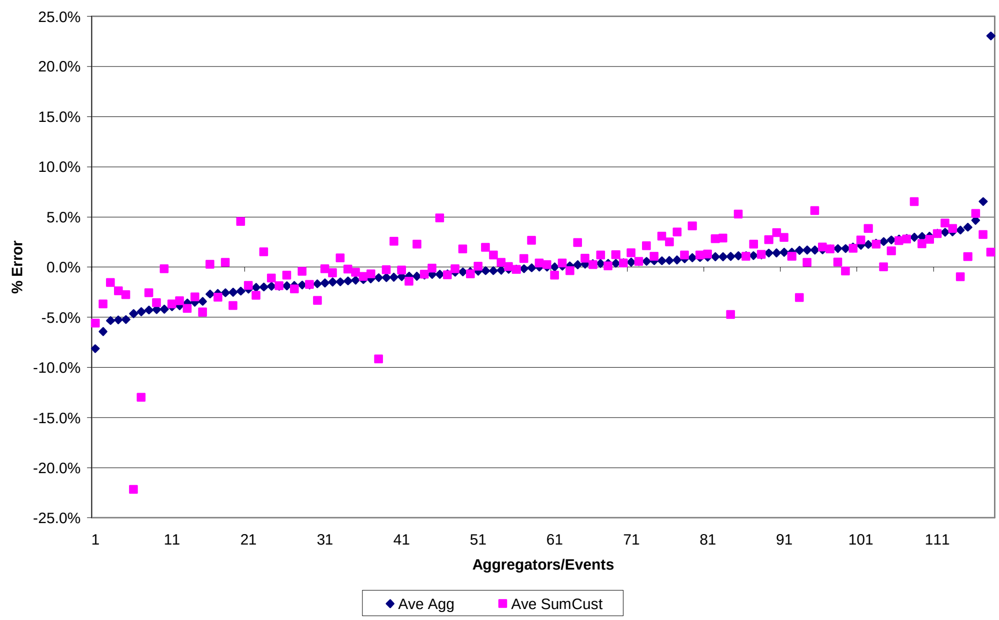


Figure 10 shows comparable values for *commercial* customer types only. In this case, the range of values is tighter and the differences between baseline-type are for the most part relatively small. Figure 11 shows values for *industrial* customer types. Here the underlying range of values is somewhat greater than for the commercial customers, and the differences between the aggregator and sum-of-customer baselines are greater. Figure 12 shows values for *schools*, which include several outliers with large errors.

[7] The percent error values across hours for a given event and aggregator tend to be similar, so that averaging errors across hours in an event simplifies the charts without discarding too much information.

Figure 10. Average Event-Day % Errors for *Adjusted 10-in-10* Aggregator and Sum-of-Customer Baselines – *Commercial*

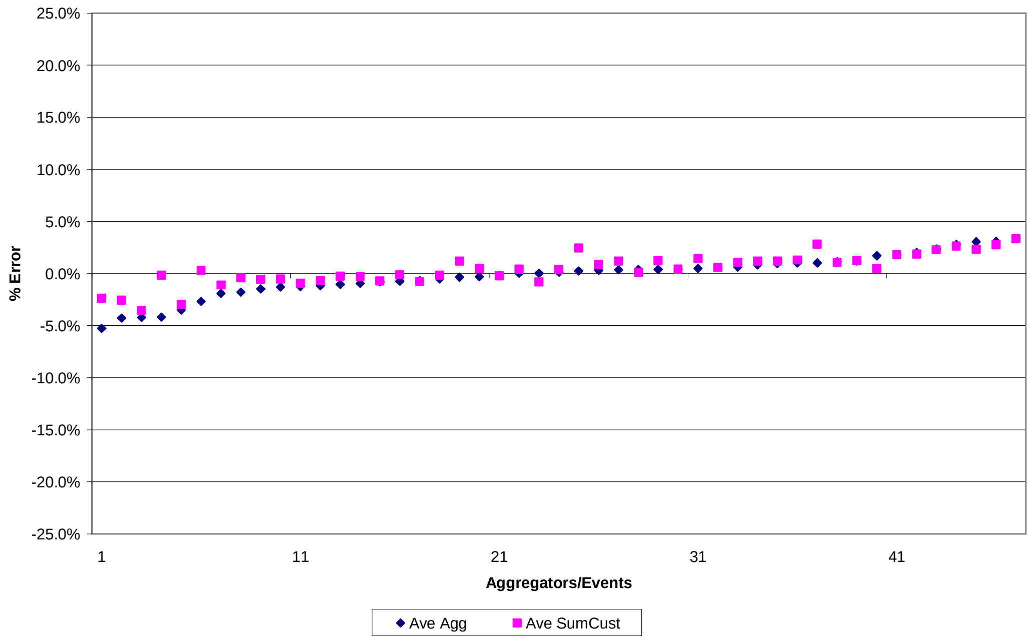


Figure 11. Average Event-Day % Errors for *Adjusted 10-in-10* Aggregator and Sum-of-Customer Baselines – *Industrial*

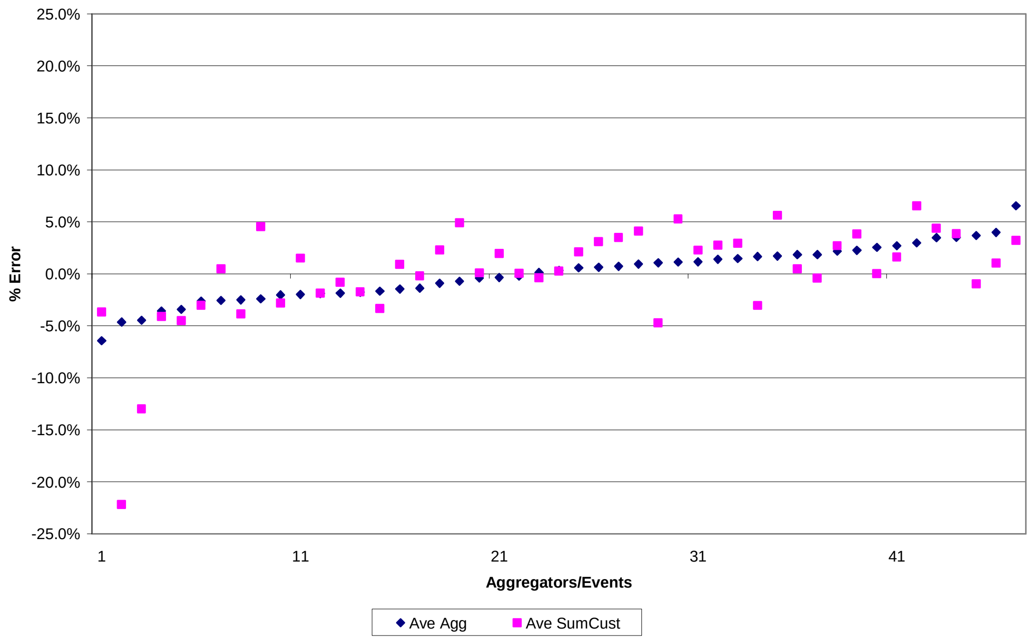
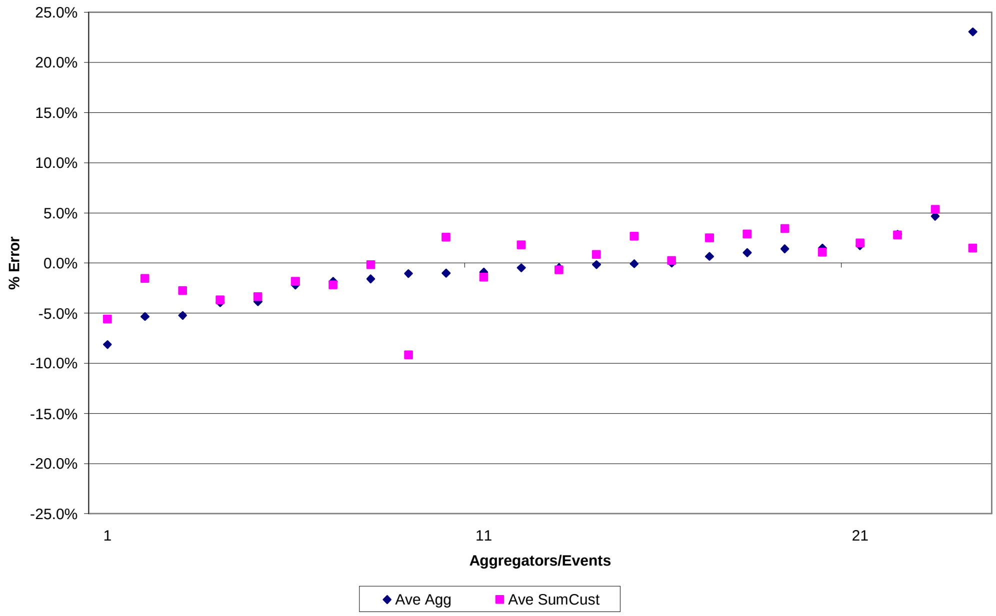


Figure 12. Average Event-Day % Errors for *Adjusted 10-in-10* Aggregator and Sum-of-Customer Baselines – *Schools*

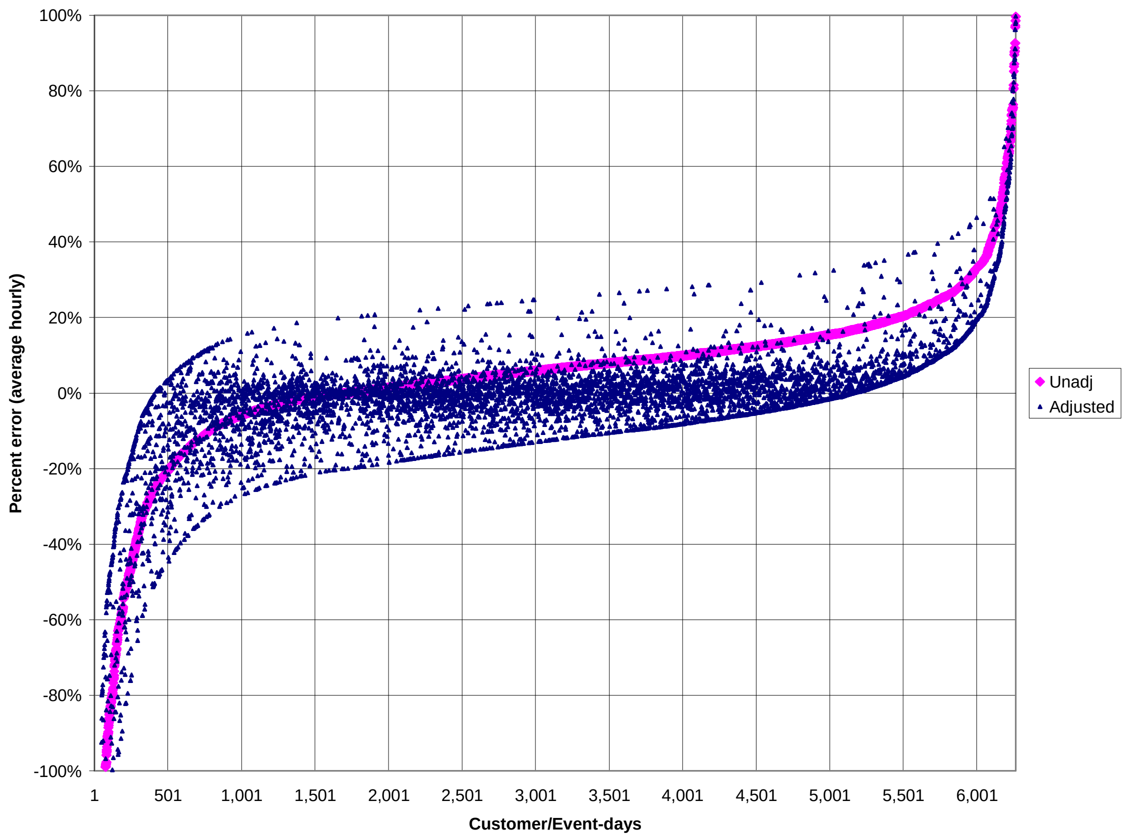


4.3.2. Customer-level distributions of baseline errors

The figures in this section are designed to illustrate the variability in relative errors at the customer level, which underlie the distributions shown in the above figures. Figure 13 shows the distribution of unadjusted and adjusted 10-in-10 baselines. The points are sorted by the values for the *unadjusted* baselines, thus providing an indication of the improvements in the percent errors due to the adjustments, as well as the breadth of the distributions across customers. 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 6.5 percent.[8] The relatively high density of *adjusted* baseline points within about 10 percent on either side of the horizontal axis indicates the extent to which the adjustments reduce the baseline errors. The resulting median percent error for the adjusted baseline is essentially zero (–0.05 percent).

[8] 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 13. Distributions of Average Event-Day % Errors for *Unadjusted* and *Adjusted 10-in-10* Baselines – *Individual Customers*



Figures 14 and 15 are designed to investigate the relative importance of the large errors that occur in at least 5 percent of cases (around 300 customers/event days) at both ends of the distribution shown in Figure 13 (*e.g.*, whether large errors tend to be associated with small or large customers). These figures plot average event-day percent errors of the adjusted 10-in-10 baseline against customer size, measured by customers' average hourly usage during event periods on non-event days, for industrial and commercial customers respectively. Figure 14 illustrates a relatively wide range of percent errors (across the horizontal axis) for industrial customers, but also demonstrates that most of the largest errors are associated with the smallest customers. The errors are also distributed reasonably symmetrically around the origin (the median of the percent errors across all customers is –0.05%). Figure 15 shows that the range of percent errors for commercial customers is tighter, with fewer extremely large errors, and the bulk of the errors are grouped fairly tightly around the origin. The largest errors are again associated with smaller customers.

Figure 14. Average Event-Day % Errors for *Adjusted 10-in-10* Baselines, by Customer Size (Average Peak kW) – *Industrial Customers*

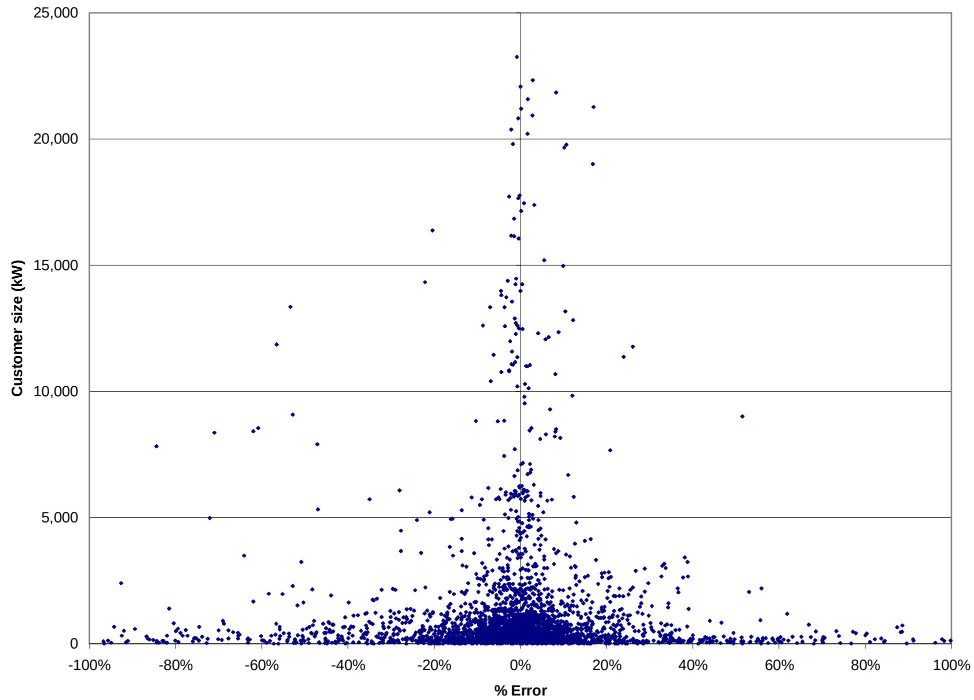
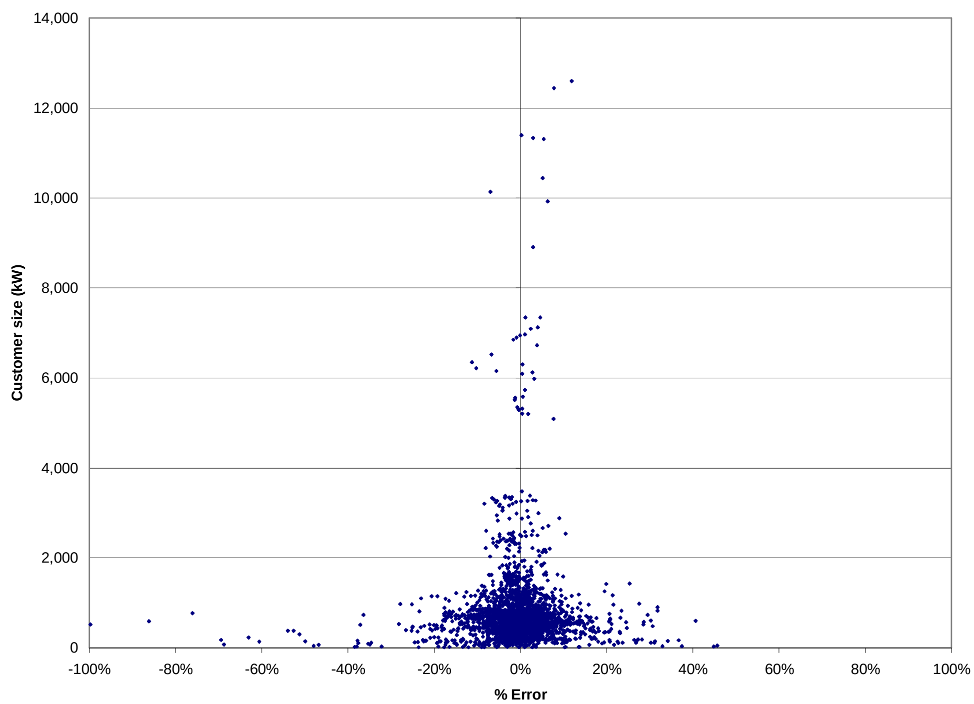


Figure 15. Average Event-Day % Errors for *Adjusted 10-in-10* Baselines, by Customer Size (Average Peak kW) – *Commercial Customers*



5. Results for ***Event Days***

This section presents baseline performance statistics for alternative baseline methods for event days, where baseline calculations are included for each aggregator that was called for each event. In this analysis, results are differentiated by both industry type and customers' choice of adjusted baseline.

5.1 Unadjusted baselines

We begin by establishing a reference point of performance results for the alternative *unadjusted* baselines on AMP event days.

5.1.1 Accuracy

Table 7 shows *accuracy* results for unadjusted versions of the three different methods based on the number of days selected for inclusion in the baseline calculation (*e.g.*, 3, 5, or 10), and for the two different methods for calculating aggregate baselines—aggregator and sum-of-customers. The following observations characterize some of the important results:

- The accuracy results for the event days are qualitatively similar to those for event-type days presented in Section 4. For the unadjusted 3-in-10 aggregator baseline, shown in the first column, and focusing on the last group of "Total" rows, relative errors average about 7 percent, for both those who selected the adjusted baseline option and those that did not.
- For the unadjusted sum-of-customer baseline, the relative errors are generally comparable, to somewhat larger than the aggregator results.
- The relative errors for the aggregator baseline generally increase with the number of days included in the baseline average.

Table 7. Accuracy of Unadjusted Baselines – Event Days
(Relative root mean square error, or Theil U-statistic)

			Aggregator			*Sum of Customers*		
			Unadjusted			**Unadjusted**		
Agg.	**Adj. BL?**	**Industry**	**3-in-10**	**5-in-10**	**10-in-10**	**3-in-10**	**5-in-10**	**10-in-10**
		Industry	0.066	0.078	0.096	0.028	0.041	0.095
		Commercial	0.110	0.143	0.180	0.103	0.135	0.180
1	No	TOTAL	0.100	0.128	0.161	0.089	0.116	0.161
		Industry	0.078	0.078	0.147	0.093	0.079	0.140
		Commercial	0.066	0.069	0.105	0.064	0.068	0.105
	No	TOTAL	0.078	0.078	0.147	0.093	0.079	0.139
		Industry	0.240	0.231	0.204	0.267	0.247	0.205
		Commercial	0.050	0.072	0.103	0.039	0.061	0.103
2	Yes	TOTAL	0.065	0.081	0.107	0.061	0.074	0.107
		Industry	0.044	0.054	0.076	0.019	0.018	0.076
		Commercial	0.044	0.065	0.092	0.022	0.042	0.090
		Schools	0.057	0.064	0.105	0.055	0.058	0.105
	No	TOTAL	0.044	0.055	0.077	0.020	0.019	0.076
		Industry	0.046	0.068	0.095	0.056	0.062	0.095
		Commercial	0.043	0.081	0.138	0.044	0.080	0.138
		Schools	0.156	0.158	0.211	0.153	0.155	0.211
3	Yes	TOTAL	0.046	0.071	0.105	0.054	0.066	0.105
		Industry	0.090	0.080	0.045	0.147	0.119	0.045
		Commercial	0.011	0.014	0.046	0.014	0.013	0.046
		Schools	n/a	n/a	n/a	n/a	n/a	n/a
	No	TOTAL	0.090	0.079	0.045	0.147	0.118	0.045
		Industry	0.095	0.089	0.096	0.142	0.118	0.096
		Commercial	0.050	0.064	0.083	0.041	0.055	0.083
		Schools	0.118	0.113	0.131	0.116	0.109	0.128
4	Yes	TOTAL	0.097	0.094	0.108	0.118	0.105	0.107
		Industry	0.064	0.067	0.074	0.086	0.071	0.073
		Commercial	0.104	0.136	0.172	0.097	0.127	0.171
		Schools	0.057	0.064	0.105	0.055	0.058	0.105
	No	TOTAL	0.073	0.083	0.097	0.088	0.083	0.097
		Industry	0.141	0.136	0.130	0.176	0.156	0.130
		Commercial	0.050	0.071	0.102	0.040	0.061	0.102
		Schools	0.118	0.113	0.131	0.116	0.109	0.128
Total	Yes	TOTAL	0.071	0.084	0.107	0.074	0.080	0.107
	All	TOTAL	0.072	0.083	0.100	0.085	0.083	0.100

5.1.2 Bias

Table 8 presents results for the bias in the unadjusted baselines. Observations include the following:

- The values in the "Total" group of rows at the bottom of the table in the first column are positive, indicating the common result that the unadjusted 3-in-10 baseline is typically *biased downward*, by 3 percent overall, and from 1 to 4 percent for different industry sub-groups.
- The sum-of-customer method produces much more variable results, with a nearly zero bias overall, and a range of upward and downward biases for various aggregators and industry types.
- Moving across the number of days included in the baseline, both the aggregator and sum-of-customer methods show increased downward biases, averaging 8 to 9 percent for the 10-in-10 method.
- The downward bias is generally larger for the commercial customer type than for the industrial.

Table 8. Bias of *Unadjusted* Baselines – Event Days
(Median percent errors)

			Aggregator			*Sum of Customers*		
			Unadjusted			Unadjusted		
Agg.	Adj. BL?	Industry	3-in-10 Un	5-in-10 Un	10-in-10 Un	3-in-10 Un	5-in-10 Un	10-in-10 Un
		Industry	7.74%	9.70%	12.04%	1.38%	5.16%	11.80%
		Commercial	11.28%	13.78%	16.93%	10.34%	13.02%	16.93%
1	No	TOTAL	9.73%	11.08%	13.71%	3.28%	6.49%	13.58%
		Industry	-2.17%	-2.51%	8.52%	-10.93%	-9.66%	8.91%
		Commercial	0.28%	1.49%	6.20%	-0.13%	1.65%	6.20%
	No	TOTAL	-1.22%	0.21%	7.96%	-5.44%	-2.88%	7.96%
		Industry	0.67%	2.99%	10.65%	-2.13%	1.92%	10.65%
		Commercial	5.76%	9.06%	11.83%	4.14%	7.61%	11.83%
2	Yes	TOTAL	3.08%	5.05%	10.65%	1.52%	3.84%	10.65%
		Industry	3.80%	4.49%	7.35%	-2.21%	0.53%	7.30%
		Commercial	4.52%	6.74%	9.46%	1.68%	4.25%	9.06%
		Schools	2.91%	4.99%	10.20%	0.91%	3.87%	10.20%
	No	TOTAL	4.13%	5.86%	7.92%	0.18%	2.64%	7.91%
		Industry	2.67%	3.19%	7.66%	-2.62%	1.44%	7.66%
		Commercial	2.11%	5.38%	10.03%	1.75%	5.13%	10.03%
		Schools	3.01%	5.61%	16.36%	2.20%	4.65%	16.36%
3	Yes	TOTAL	2.67%	3.19%	8.96%	-1.13%	1.44%	8.96%
		Industry	-7.85%	-6.88%	-3.68%	-13.60%	-10.34%	-3.70%
		Commercial	-0.49%	0.68%	4.14%	-1.05%	0.70%	4.14%
		Schools	n/a	n/a	n/a	n/a	n/a	n/a
	No	TOTAL	-3.21%	-1.99%	1.20%	-11.38%	-8.42%	1.13%
		Industry	-0.55%	1.04%	5.04%	-6.33%	-2.55%	5.04%
		Commercial	5.00%	6.41%	8.24%	3.79%	5.42%	8.24%
		Schools	0.15%	2.04%	7.91%	-1.11%	0.88%	7.29%
4	Yes	TOTAL	2.93%	3.86%	7.72%	-0.25%	1.96%	7.72%
		Industry	2.56%	4.73%	7.88%	-4.27%	0.39%	7.87%
		Commercial	3.82%	5.69%	7.78%	1.82%	3.71%	7.78%
		Schools	2.91%	4.99%	10.20%	0.91%	3.87%	10.20%
	No	TOTAL	3.55%	5.31%	8.00%	-0.56%	2.36%	8.00%
		Industry	0.99%	2.92%	9.48%	-4.46%	0.03%	9.48%
		Commercial	4.38%	6.67%	9.00%	3.23%	5.55%	9.00%
		Schools	0.15%	2.04%	11.20%	-1.11%	0.88%	10.59%
Total	Yes	TOTAL	2.90%	4.49%	9.48%	0.98%	3.37%	9.48%
	All	TOTAL	3.11%	4.69%	8.47%	-0.20%	2.82%	8.70%

5.2 Adjusted baselines

This section shows accuracy and bias results for *adjusted* versions of each of the alternative baseline methods. Note that the customers who did *not* select the adjusted baseline option actually faced the unadjusted 3-in-10 baseline in the AMP events, while those who did select the adjusted baseline faced the adjusted 3-in-10 baseline.

5.2.1 Accuracy

Table 9 shows *accuracy* results for the various adjusted versions of the two methods for aggregating customers. Key findings include the following:

- Focusing first on the first column in each group of columns, for the adjusted 3-in-10 baseline, and the bottom sets of rows showing results for all customers, the adjusted baseline shows smaller relative errors than the corresponding unadjusted baseline in nearly every case, with an overall relative error of 2.6 percent compared to 7.2 percent for the unadjusted version.

- The smaller relative errors hold for every customer sub-group.
- For the aggregator baselines in particular, the relative errors of the adjusted baselines are very similar across the number of days included in the baseline.
- Similar results are obtained for the sum-of-customers method, though the relative errors are not reduced as much compared to the unadjusted baseline as for the aggregator method.
- For those who chose the adjusted baseline, accuracies improve across the number of days included in the baseline, for both the aggregator and sum-of-customers method.
- For those who remained on the unadjusted baseline, the accuracy results are more varied; accuracy would generally improve with adjustment, but there is not a clear winner in terms of number of days in the baseline average.
- The adjusted baselines are generally more accurate for commercial than for industrial types, while schools generally show the least accuracy.

Table 9. Accuracy of *Adjusted* Baselines – Event Days
(Relative root mean square error, or Theil U-statistic)

			Aggregator			*Sum of Customers*		
			Symmetric Adjustment			Symmetric Adjustment		
Agg.	**Adj. BL?**	**Industry**	**3-in-10**	**5-in-10**	**10-in-10**	**3-in-10**	**5-in-10**	**10-in-10**
		Industry	0.018	0.013	0.017	0.028	0.014	0.036
		Commercial	0.014	0.023	0.058	0.017	0.028	0.063
1	No	TOTAL	0.015	0.021	0.050	0.021	0.025	0.057
		Industry	0.048	0.039	0.043	0.070	0.050	0.062
		Commercial	0.116	0.100	0.061	0.088	0.104	0.055
	No	TOTAL	0.048	0.039	0.043	0.070	0.050	0.062
		Industry	0.105	0.097	0.052	0.180	0.167	0.113
		Commercial	0.032	0.023	0.020	0.029	0.023	0.025
2	Yes	TOTAL	0.037	0.028	0.022	0.043	0.037	0.032
		Industry	0.016	0.018	0.015	0.009	0.009	0.031
		Commercial	0.017	0.015	0.013	0.033	0.025	0.011
		Schools	0.046	0.036	0.022	0.045	0.033	0.017
	No	TOTAL	0.017	0.018	0.015	0.011	0.010	0.030
		Industry	0.034	0.036	0.026	0.055	0.041	0.028
		Commercial	0.031	0.043	0.013	0.031	0.043	0.014
		Schools	0.100	0.075	0.098	0.108	0.078	0.095
3	Yes	TOTAL	0.034	0.038	0.025	0.052	0.042	0.027
		Industry	0.027	0.022	0.019	0.102	0.089	0.046
		Commercial	0.047	0.038	0.037	0.042	0.041	0.038
		Schools	n/a	n/a	n/a	n/a	n/a	n/a
	No	TOTAL	0.028	0.022	0.019	0.101	0.089	0.046
		Industry	0.045	0.039	0.048	0.076	0.068	0.061
		Commercial	0.014	0.017	0.021	0.010	0.015	0.028
		Schools	0.062	0.078	0.083	0.064	0.077	0.079
4	Yes	TOTAL	0.048	0.054	0.060	0.063	0.065	0.064
		Industry	0.022	0.022	0.026	0.059	0.053	0.042
		Commercial	0.015	0.022	0.055	0.020	0.028	0.059
		Schools	0.046	0.036	0.022	0.045	0.033	0.017
	No	TOTAL	0.021	0.022	0.033	0.054	0.049	0.045
		Industry	0.064	0.058	0.048	0.109	0.099	0.075
		Commercial	0.032	0.023	0.020	0.029	0.023	0.025
		Schools	0.062	0.078	0.083	0.064	0.077	0.079
Total	Yes	TOTAL	0.039	0.034	0.032	0.047	0.043	0.039
	All	TOTAL	0.026	0.026	0.032	0.053	0.048	0.044

5.2.2 Bias

Table 10 shows *bias* results for the various adjusted versions of the two methods for aggregating customers. Key findings include the following:

- The first column in each of the two groups of columns, for the adjusted 3-in-10 baseline, generally contain negative numbers across all aggregators, and in total, implying *upward biases* for nearly every subgroup, ranging from near-zero to two outliers near 9 and 10 percent for sum-of-customer baselines.
- For the customers selecting the adjusted baseline, in the last set of rows at the Total level for the Aggregator baseline, the bias is nearly zero for Commercial, and an upward bias of 1.8 percent for Industrial, which compare to 4.4 and 1 percent *downward* biases for the *unadjusted* 3-in-10.
- For the customers *not* selecting the adjusted baseline, the adjusted 3-in-10 reduces the median % error in the unadjusted baseline from a nearly 3 percent understatement to nearly zero for Industrial customers, and from a nearly 4 percent understatement for Commercial customers to a typical upward bias of 2 percent.
- The 3-in-10 sum-of-customers baseline shows similar median % errors that are slightly larger in absolute value than for the aggregator method.
- The results in the third column, for the adjusted 10-in10 baseline, suggest the smallest biases overall.

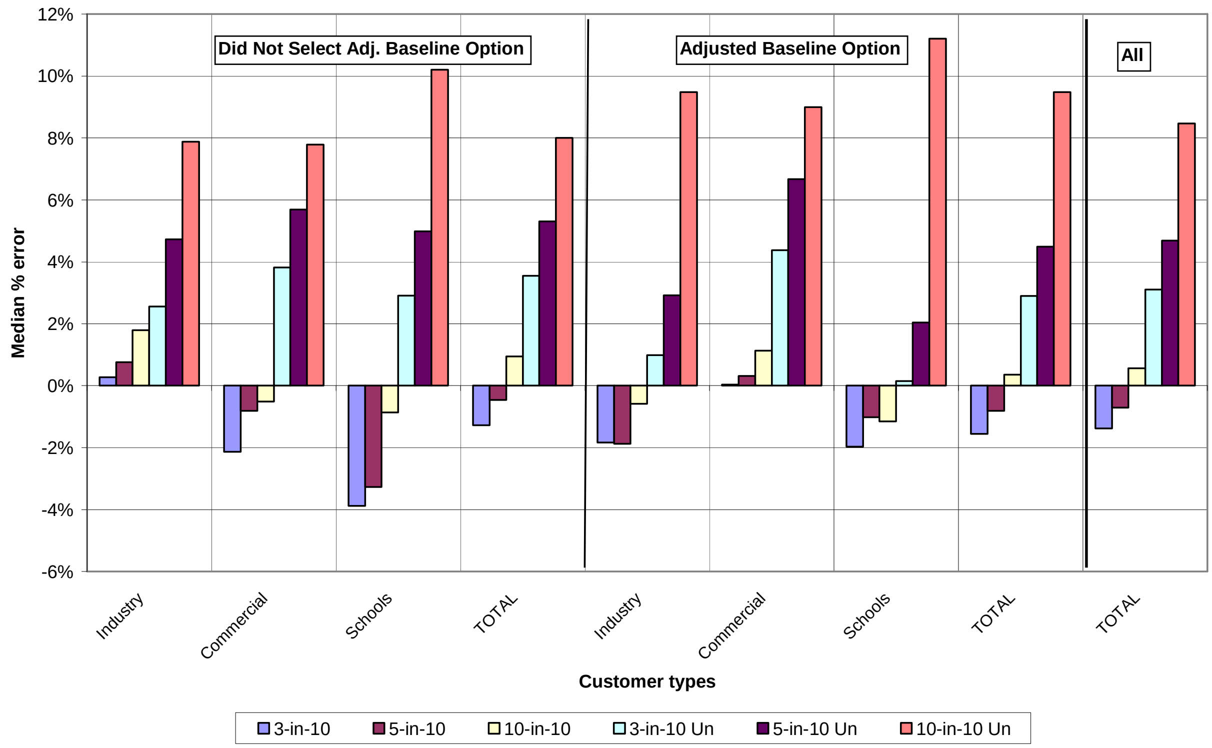
Table 10. Bias of *Adjusted* Baselines – Event Days
(Median percent errors)

			Aggregator			*Sum of Customers*		
			Symmetric Adjustment			Symmetric Adjustment		
Agg.	Adj. BL?	Industry	3-in-10	5-in-10	10-in-10	3-in-10	5-in-10	10-in-10
1	No	Industry	1.11%	1.32%	1.58%	-1.87%	-0.09%	4.41%
		Commercial	0.37%	1.03%	0.95%	0.74%	1.35%	3.43%
		TOTAL	0.53%	1.11%	1.30%	-1.22%	0.71%	3.64%
2	No	Industry	-2.47%	-3.99%	2.79%	-8.62%	-4.50%	2.54%
		Commercial	-6.35%	-6.06%	-1.00%	-5.46%	-6.25%	-1.08%
		TOTAL	-4.73%	-4.71%	1.78%	-6.29%	-4.81%	1.09%
	Yes	Industry	-4.02%	-4.81%	-3.51%	-2.99%	-1.83%	0.77%
		Commercial	-2.10%	-1.57%	0.08%	-1.38%	-0.06%	2.06%
		TOTAL	-3.00%	-2.76%	-0.73%	-2.90%	-1.19%	0.95%
3	No	Industry	1.19%	1.64%	1.67%	-0.22%	0.15%	2.75%
		Commercial	-1.24%	-1.03%	-1.11%	-3.19%	-2.51%	-0.02%
		Schools	-3.88%	-3.28%	-0.86%	-3.84%	-2.67%	1.30%
		TOTAL	-1.00%	-0.81%	0.28%	-2.36%	-1.42%	2.05%
	Yes	Industry	-3.78%	-1.87%	-0.55%	-5.02%	-3.44%	-0.24%
		Commercial	-3.19%	-3.23%	0.51%	-3.20%	-3.12%	0.64%
		Schools	-7.23%	-4.09%	3.65%	-7.13%	-4.75%	4.82%
		TOTAL	-3.63%	-1.87%	0.07%	-3.77%	-3.44%	0.27%
4	No	Industry	-2.13%	-1.20%	-1.42%	-9.90%	-8.58%	-3.93%
		Commercial	-4.45%	-2.61%	-3.15%	-3.24%	-3.39%	-3.17%
		Schools	n/a	n/a	n/a	n/a	n/a	n/a
		TOTAL	-2.84%	-1.53%	-2.15%	-6.65%	-5.94%	-3.17%
	Yes	Industry	-0.14%	-0.37%	0.14%	-2.22%	-0.71%	1.91%
		Commercial	0.64%	1.36%	1.96%	0.22%	1.33%	2.35%
		Schools	-1.90%	-1.02%	-1.15%	-3.64%	-1.76%	-0.37%
		TOTAL	0.03%	0.18%	1.27%	-0.71%	0.11%	1.40%
Total	No	Industry	0.28%	0.76%	1.79%	-2.64%	-0.58%	3.31%
		Commercial	-2.13%	-0.81%	-0.51%	-2.69%	-2.20%	-0.38%
		Schools	-3.88%	-3.28%	-0.86%	-3.84%	-2.67%	1.30%
		TOTAL	-1.28%	-0.46%	0.94%	-2.74%	-1.65%	2.15%
	Yes	Industry	-1.83%	-1.87%	-0.58%	-3.36%	-1.97%	0.56%
		Commercial	0.04%	0.32%	1.13%	-0.22%	0.39%	1.61%
		Schools	-1.97%	-1.02%	-1.15%	-3.72%	-1.76%	-0.19%
		TOTAL	-1.56%	-0.81%	0.36%	-2.71%	-1.02%	1.10%
	All	TOTAL	-1.38%	-0.71%	0.57%	-2.74%	-1.19%	1.38%

Figure 16 provides a comparison of the results in the lower "Total" panel of Table 10 for the Aggregator method to the comparable results in Table 8 for unadjusted baseline. Each set of bars shows the median % errors for the three adjusted baselines and then the three unadjusted baselines. The first set of panels presents results for those customers who did not select the adjusted baseline option, while the second set of panels shows results for those who did select the option. The final set of bars shows results for all customers.

The figure clearly shows the typical result that the downward bias (positive median % error) of the unadjusted baseline becomes greater as the number of days included in the baseline average expands, with the largest bias for the 10-in-10 baseline. The figure also clearly shows the smaller biases of the adjusted baselines, with the adjusted 10-in-10 often producing the smallest bias. For the commercial customers who might be most interested in the adjusted baseline, the bias of the adjusted baseline appears smallest, though still understating the true baseline by a small amount.

Figure 16. Comparison of Bias of *Adjusted* and *Unadjusted* Baselines – *Aggregator Method*



5.3 Conclusions for event days

The performance of the alternative baseline methods on event days, in terms of accuracy and bias, appears qualitatively similar to their performance on the *event-type* days presented in Section 4. The aggregator method appears to do better than the sum-of-customer method. Adjusting the baseline for morning usage generally improves the accuracy and reduces the bias of the unadjusted baselines. Performance results vary considerably across aggregators and industry types. The adjusted 10-in-10 does not dominate the other methods as it appeared to do for the event-type days. However, it performs at least as well and often better than the other adjusted baselines.

6. Gaming

An issue of concern for adopting the adjusted baseline method is whether customers and aggregators would try or succeed in "gaming" the baseline by artificially increasing usage in the morning hours that are used to construct the adjustment factor. Such an increase could have the effect of increasing the baseline used for settlement, and hence the achieved load impacts on event days.

We looked for evidence of gaming among the aggregators who offered an adjusted baseline option and the customers who accepted it. We examined the issue from two directions. First, we constructed aggregate load profiles for all of the AMP event and event-type days for each aggregator, by industry type and choice of adjusted baseline. We then examined the event-day loads for evidence of increases in usage prior to the events compared to typical usage patterns in the same hours on event-type days. Second, for all AMP customers we calculated average usage in pre-event hours on both event days and event-type days, and examined the ratios of the two values for evidence of significantly higher values on event days. Results of both methods are summarized below.

6.1 Comparison of loads on event-days and event-type days

The following charts show aggregated loads for sub-groups of customers (*e.g.*, by industry type and choice of adjusted baseline) for aggregators that offered an adjusted baseline option, and for which a reasonable number of customers selected the option. The load profiles are shown for the five event days (only some aggregators were called for some of the events) and the ten event-type days that were used in the baseline analysis. Bold lines indicate days on which events were called for that aggregator. For some aggregators and sub-groups, the loads appear to be grouped at two different usage levels. This typically occurs due to customers being added to or removed from the group during the summer period, through changes in monthly nominations. Graphs are shown primarily for sub-groups that selected the adjusted baseline. In a few cases, graphs are also shown for groups that did not select the adjusted baselines.

Looking across the figures, the load reductions on event days are usually quite evident, with the industrial customers typically showing the largest reductions during the event hours. Across the aggregators and sub-groups, there is only one instance of an event-day load that takes on a shape potentially indicative of gaming. That instance is shown in Figure 18, for the industrial sub-group of the second aggregator. This is a relatively small group, with peak-period demand of about 15 to 16 MW. On further investigation, the group is dominated by one large customer who joined mid-way through the summer, which explains the two different typical load profile levels for the group prior to and after August 1. Examination of that customer’s loads indicates somewhat variable loads, like many industrial customers, on some days operating at levels that are half that on other days. On the event day in question, September 5, the customer’s load began at a level suggestive of a lower level of operations, particularly following a similar pattern as the previous day. However, around noon the load increases by about 4 MW and stays there until the hour prior to the event, at which time it drops by about 5 MW.

It is not possible to know with certainty whether this load profile is indicative of actual gaming behavior. However, it is at least suggestive of how such gaming behavior could be conducted.

Figure 17. Aggregator 2; Industrial; *No Adjusted BL*

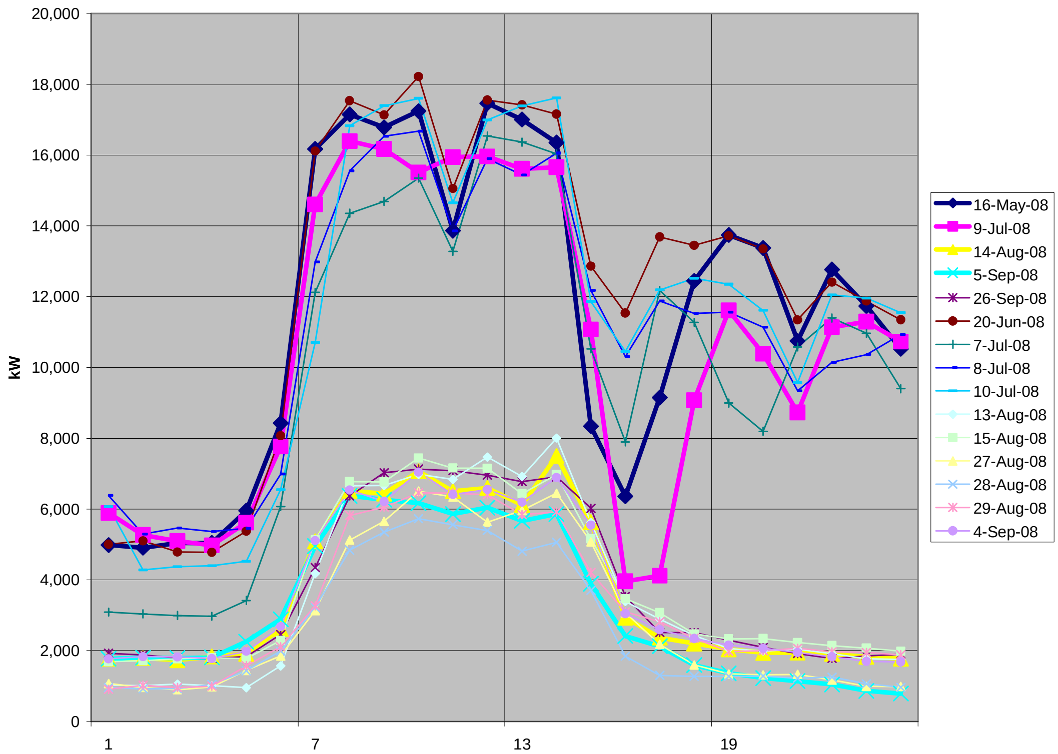


Figure 18. Aggregator 2; Industrial; *Adjusted BL Option*

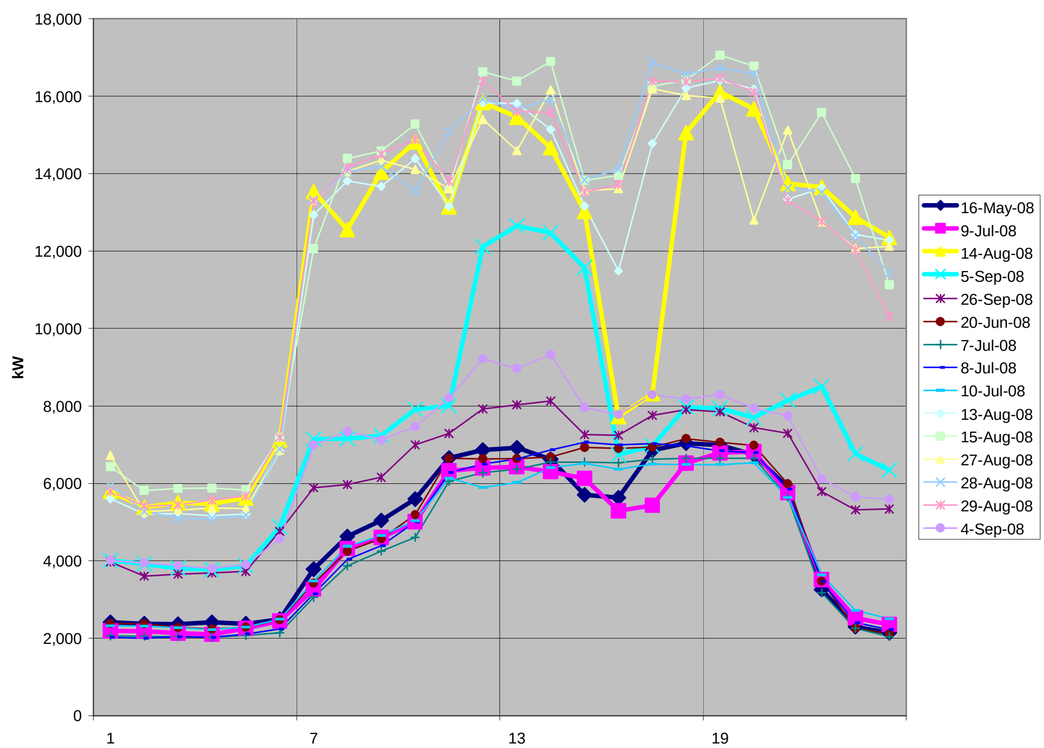


Figure 19. Aggregator 2; Commercial; *Adjusted BL Option*

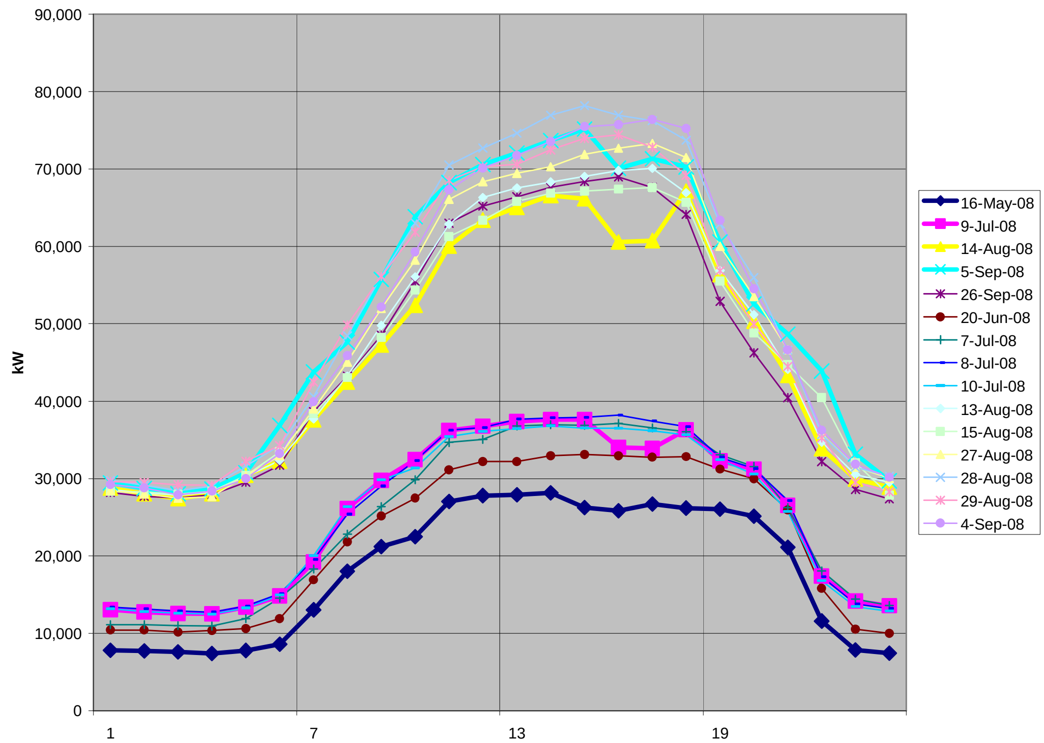


Figure 20. Aggregator 3; Industrial; *Adjusted BL Option*

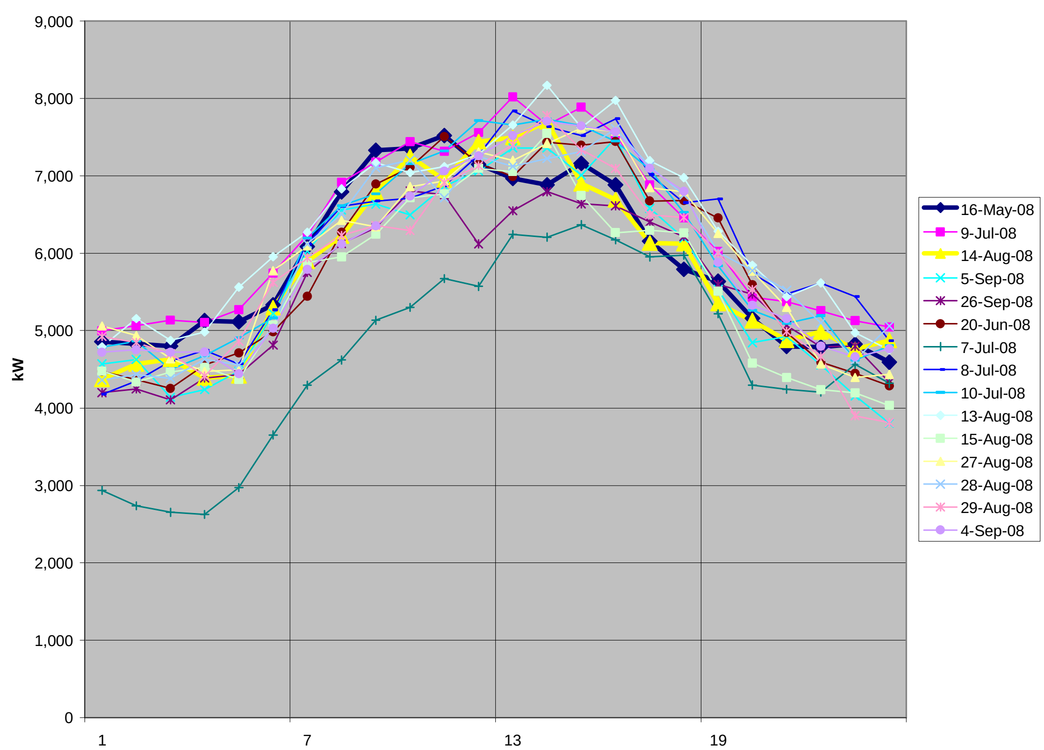


Figure 21. Aggregator 4; Industrial; *No Adjusted BL*

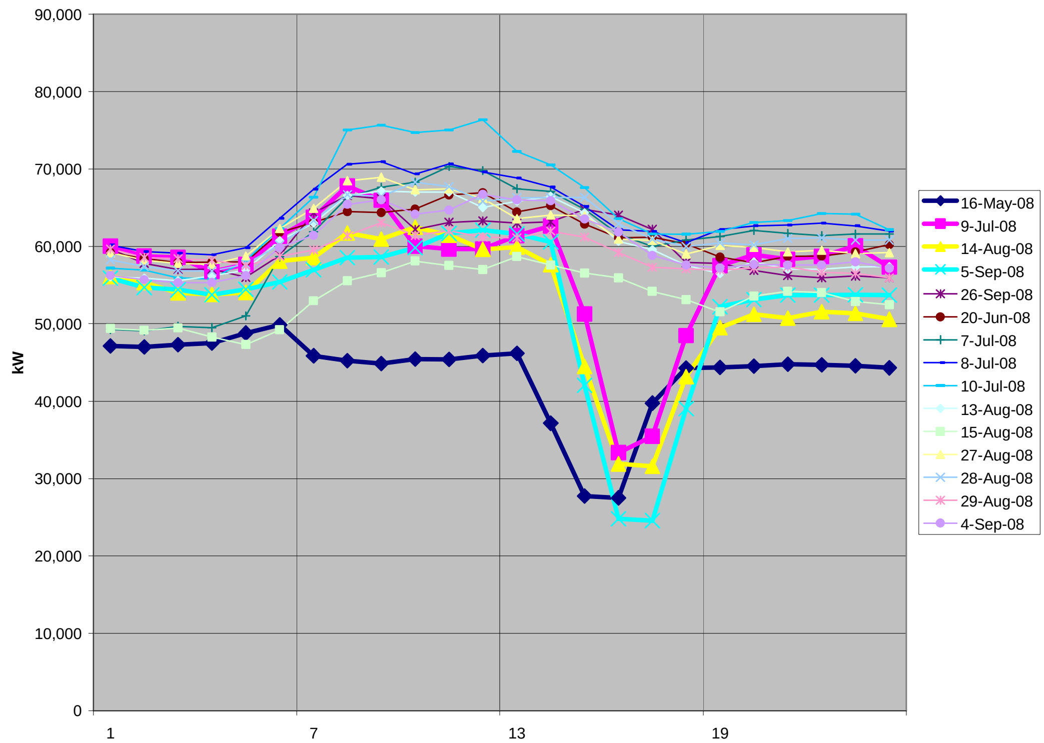


Figure 22. Aggregator 4; Industrial; *Adjusted BL Option*

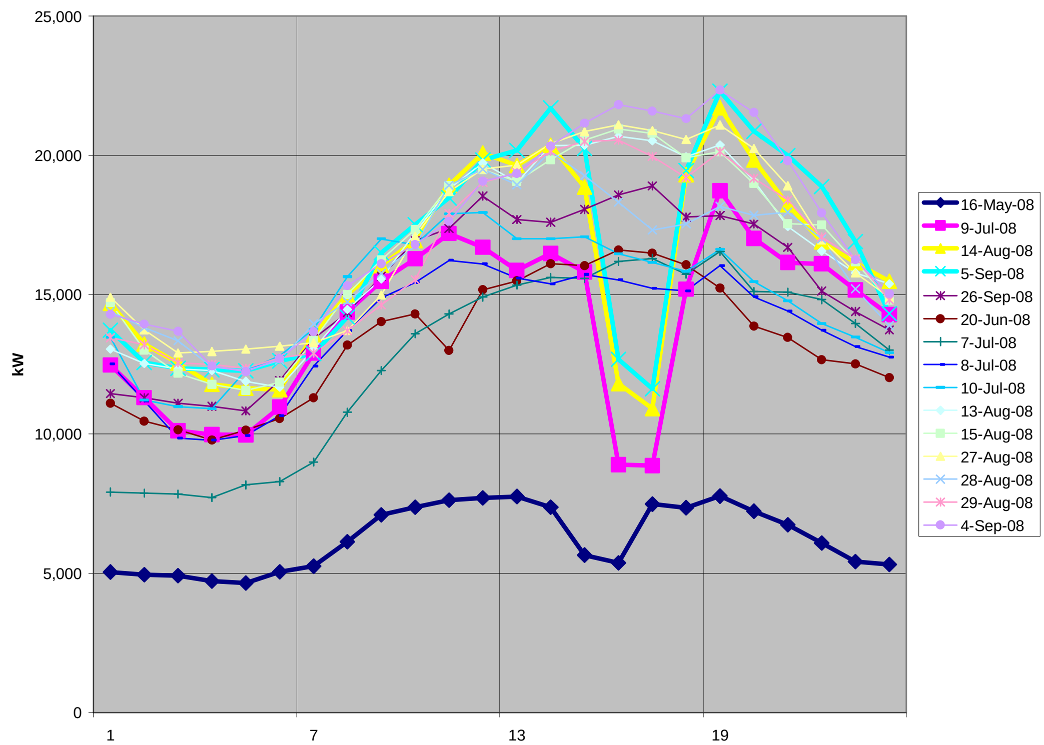


Figure 23. Aggregator 4; Commercial; *Adjusted BL Option*

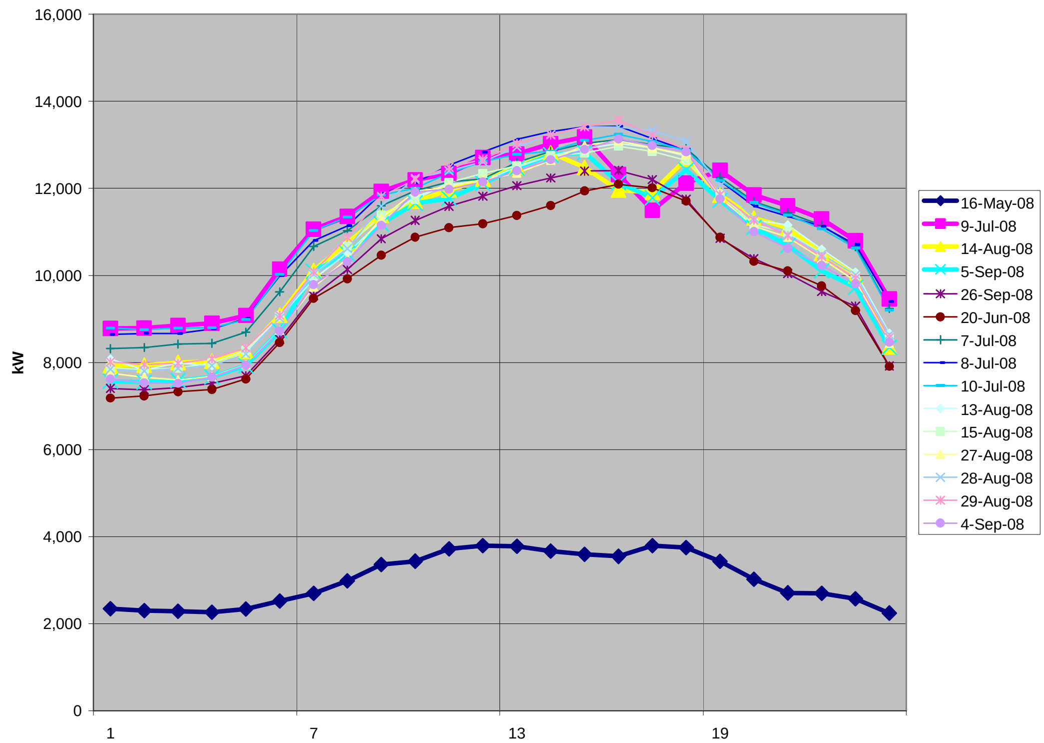
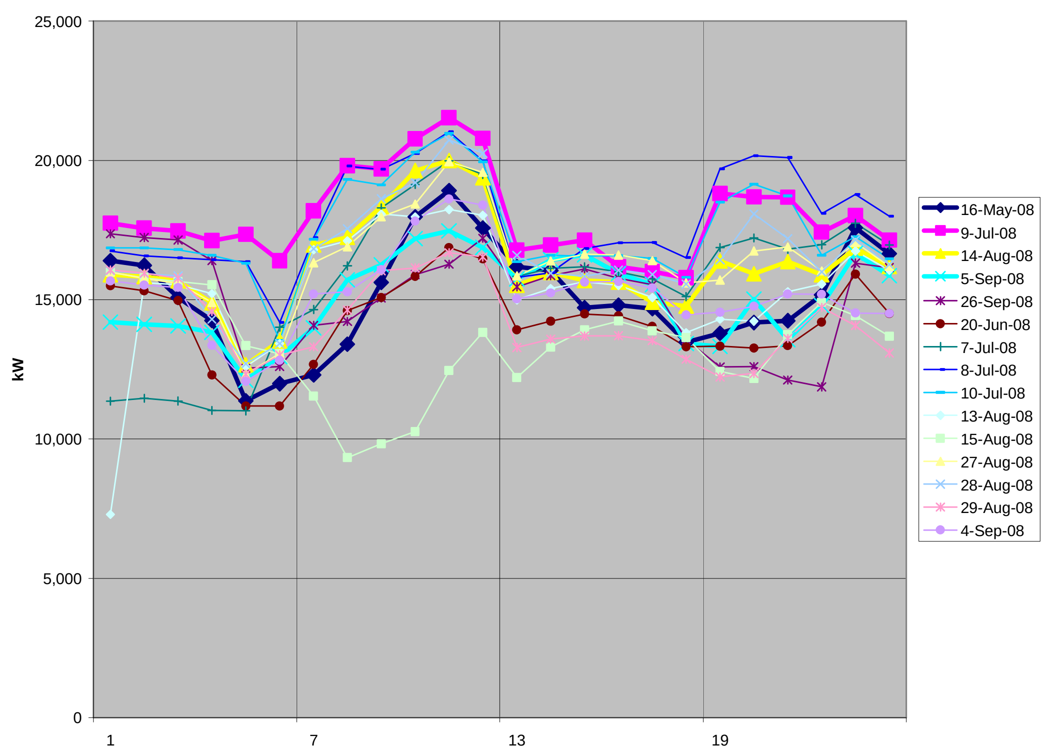


Figure 24. Aggregator 4; Schools; *Adjusted BL Option*



6.2 Analysis of pre-event usage

This section presents results of our analysis of AMP customers' typical pre-event usage levels on event days compared to that usage level on event-type days that were not called as events. Table 11 summarizes the number of customers, and the averages, standard deviations and coefficients of variation of ratios of their pre-event usage levels across customers in the three customer types, and by their choice of the adjusted baseline option. As seen in the second set of columns, the average values of the event-day to event-type day pre-event usage ratio are near 1.0, and differ only negligibly between those that accepted the adjusted baseline option and those that did not.

Table 11. Ratios of Average Morning Usage – Event and Non-Event Days (*By Choice of Adjusted Baseline*)

Customer type	Count		Ave. AM kWh - Event/ Non-event		SD		CV	
	No	Yes	No	Yes	No	Yes	No	Yes
1. Ind	191	47	1.05	1.07	1.09	0.25	1.04	0.23
2. Comm'l	92	106	0.99	0.99	0.05	0.18	0.05	0.18
3. Schools	9	6	1.01	1.00	0.17	0.11	0.16	0.11
Grand Total	292	159	1.03	1.01	0.88	0.20	0.86	0.20

Table 12 provides additional detail on the distributions of the ratios, showing quartile values. The median values, like the averages in Table 11, are close to one, for both those who selected adjusted baselines and those who didn't. Differences in the distributions of values are also modest, except that the spread of the non-adjusted industrial group is broader than that for the adjusted baseline group. None of these values suggest concern about systematic gaming efforts.

Table 12. Quartiles of Ratios of Average Morning Usage

Quartiles	Not Adjusted		Adjusted BL	
	Ind	Comm'l	Ind	Comm'l
Min	0.04	0.85	0.55	0.59
First quartile	0.85	0.97	0.98	0.94
Median	0.99	0.99	1.04	0.98
Third quartile	1.07	1.02	1.10	1.02
Max	3.17	1.22	2.09	2.61

6.3 Conclusions about gaming

The analysis of sub-group level aggregated load data and individual customer pre-event usage on event days and event-type days finds little if any evidence of artificial increases in pre-event usage in an attempt to "game" the adjusted baseline. The sub-group load profiles show little difference between groups that faced adjusted baselines and those that did not, and the load profiles for adjusted-baseline groups show little difference between event days and event-type non-event days. Only one case was found, for one industrial customer of one aggregator, in which the load rose unusually in the four hours prior to one event, possibly indicating an attempt to increase the baseline from which the load impact would be measured.

Analysis of the distribution of ratios of pre-event usage on event and event-type days confirmed the findings from the aggregated load data, revealing no evidence of systematic increases in pre-event consumption on event days.

7. Conclusions

7.1 Baseline performance – event-type days

The results of this baseline analysis provide a reasonably consistent story regarding the baseline issues of the relative accuracy of aggregator and sum-of-customer 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 accuracy of the *aggregator* method of calculating baselines compared to the *sum-of-customer* method, the results suggest that the aggregator method is more accurate, but not by a wide margin.
    a. Accuracy results for the *unadjusted* versions of the two methods suggested relative errors ranging from 5 to 8 percent. Results were similar for three of four aggregators, but the aggregator method was substantially better for one of them, leading to somewhat more accurate aggregator results overall.
    b. Accuracy statistics for *adjusted* versions of both baseline methods improved substantially over the unadjusted baselines for both methods, reducing relative errors by about half. Accuracy results were somewhat better for the aggregator method.
    c. For the most accurate adjusted method (adjusted 10-in-10), accuracy results were quite similar for both aggregator and sum-of-customer baselines.
    d. The accuracy of the two methods varied by industry type as well as aggregator, with accuracy generally better for commercial customers than for industrial customers or schools.
2. 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:
    a. The unadjusted 3-in-10 baselines suggest a typical *downward bias* of around 4 percent at the aggregator level, rising to 5 or 6 percent for commercial customers.
    b. The morning adjustments generally convert the typical downward bias (under-statement) to a small upward bias of less than one percent, as measured by the median percent error.
    c. The morning adjustments for the sum-of-customer baselines generally increase the upward bias by more than for the aggregator baselines.
3. 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.
    a. The relative accuracy of adjusted versions of all three types of the *aggregator* baseline produced very similar results, with relative errors ranging from 2 to 3 percent.
    b. For the *sum-of-customers* baseline method, the accuracy of the adjusted baselines improved somewhat moving from the 3-day to 10-day method.

    c. The *bias* results for both the aggregator and sum-of-customers method generally improved when moving from the 3-, to 5-, to 10-day method, with the improvement being greatest for the sum-of-customers method.
4. Examination of the performance of *upward-only* adjustments to the 5-in-10 and 10-in-10 baseline methods suggests that they reduce baseline performance, but not dramatically.
    a. The two upward-only adjustments reduced the accuracy of the aggregator baseline only slightly compared to the symmetric adjustments, but reduced the accuracy of the sum-of-customers baseline more substantially for some aggregators and industry types.
    b. The upward-only adjustments increased the bias of the aggregator baseline modestly, particularly for the fourth aggregator, but increased the bias more substantially for the sum-of-customers baseline, not unexpectedly (due to the upward-only adjustments) producing greater *upward bias*.
5. Comparing unadjusted 5-in-10 and 10-in-10 baselines to comparable symmetric adjusted versions illustrates the improved performance of the adjusted versions, which should be taken into account in any decision to allow a choice among those options.
    a. The adjusted 5-in-10 and 10-in-10 baselines are substantially more accurate than unadjusted versions, with relative errors approximately half that of unadjusted versions.
    b. The adjusted versions also have smaller biases, whereas the unadjusted versions have median relative errors suggesting typical understated baselines of 5 percent or more.
6. Examination of the variability of percent errors of 10-in-10 baselines for *individual customers* illustrates the likely source of greater baseline errors in sum-of-customer baselines compared to aggregator baselines.
    a. Morning adjustments improve the accuracy and reduce the bias of the unadjusted 10-in-10 baselines at the individual customer level.
    b. The *range* of errors is greater for industrial customers than for commercial customers, with a number of large over-stated baselines (*e.g.*, approximately 5 percent of customer/events have average percent errors exceeding 40 percent).
    c. Examination of the relationship between the magnitudes of relative errors and customer size suggests that the greatest errors are generally associated with the smallest customers.

7.2 Baseline performance – event days

The performance of the alternative baseline methods on event days, in terms of accuracy and bias, appears qualitatively similar to their performance on the *event-type* days summarized above. The aggregator method appears to do better than the sum-of-customer method. Adjusting the baseline for morning usage generally improves the accuracy and reduces the bias of the unadjusted baselines. Performance results vary considerably across aggregators and industry types. The adjusted 10-in-10 does not dominate the other methods as it appeared to do for the event-type days. However, it performs at least as well and often better than the other adjusted baselines.

7.3 Was gaming successfully avoided?

Analysis in this study revealed no evidence of systematic increases in pre-event consumption on event days that would be indicative of attempts to game the adjusted baseline. Only one case was found, for one industrial customer of one aggregator, in which hourly usage rose unusually in the four hours prior to one event, possibly indicating an attempt to increase the baseline from which the load impact would be measured.

The evidence in this baseline analysis suggests that adjusted baselines are more accurate and less biased than unadjusted baselines. However, widespread adoption of adjusted baselines would seem to call for monitoring, possibly during the event season, to check for unusual load changes that could indicate gaming behavior. Creation and examination of aggregator load profiles like those examined in Section 6 could serve as an example.

Appendix: Technical Background

It may be useful to point out several related features of any baseline analysis that involves comparisons of alternative baseline methods. These include the relationships between baseline definitions, baseline errors, and implied differences in estimated load impacts. The present baseline analysis differs from previous analyses due to the additional objective of measuring baselines for *aggregated* groups of customers.

Baseline definitions

Consider the following definitions:

Individual baseline: $IBL^i_d = f(E^i_{d-t})$,

Aggregator baseline: $ABL_d = f(\sum E^i_{d-t})$,

Sum of customer baselines: $SBL_d = \sum IBL^i_d = \sum f(E^i_{d-t})$.

For simplicity, assume that the baselines are calculated as the average across hours in an event. Thus, the value *E* represent average hourly load during the event period, the superscript, *i*, refers to an individual customer, *d* refers to the event day, and the function *f* refers to a rule for calculating the baseline across previous days, (*d–t*) (*e.g.*, average of highest 3 days in previous 10 eligible days). The *aggregator* baseline applies the baseline definition to the aggregated load of customers in the group, while the *sum of customer* baseline adds up the calculated baselines of each individual customer in the group.

Baseline errors

Baseline analyses typically calculate and compare different measures of baseline errors, defined as the difference between the *true* baseline (TBL) and the estimated baseline, as defined above. For example, baseline errors for an individual customer and an aggregated group of customers may be written as:

$$ERRI^i_d = TBL^i_d - IBL^i_d \text{ , and}$$

$$ERRA_d = \sum TBL^i_d - ABL_d.$$

When dealing with event-type days on which events were not actually called, the true baseline equals actual consumption during the "event" period. Given the interest in comparing the performance of the aggregator and sum-of-customer baselines, we can define the difference in errors for those two baselines as:

$$\begin{aligned} DiffERR &= ERRA_d - \sum ERRI^i_d \,. \\ &= \sum TBL^i_d - ABL_d - (\sum TBL^i_d - \sum IBL^i_d) \\ &= - ABL_d + \sum IBL^i_d. \end{aligned}$$

That is, differences in the *errors* of the two baselines are equal to the differences between the two baselines (*i.e.*, for purposes of comparing the errors of two alternative baselines, the true baselines drop out of the consideration). In the current baseline analysis, the primary interest is in differences in the accuracy and bias of different baseline methods, both of which statistics are functions of baseline errors across a number of events, and customers or aggregators. However, the performance statistics for each baseline method are of interest in themselves, so that we calculate the baseline errors relative to the true baseline and then compare results.

Differences in load impacts and baseline errors

Load impacts (*i.e.*, differences between the baseline and actual load) corresponding to the alternative baseline methods may be written as follows:

Individual load impact (ILI): $IBL^{i}_{d} - E^{i}_{d}$,

Aggregate load impact (ALI): $ABL_{d} - \sum E^{i}_{d}$,

Sum of customer load impact (SLI): $SBL_{d} - \sum E^{i}_{d}$.

The difference between the aggregator load impact and the sum of customer load impacts may be written as:

$$\begin{aligned}
\text{DiffLI} &= \text{ALI} - \text{SLI} \\
&= (ABL_{d} - \sum E^{i}_{d}) - (SBL_{d} - \sum E^{i}_{d}) \\
&= ABL_{d} - SBL_{d} \\
&= ABL_{d} - \sum IBL^{i}_{d} \\
&= -\text{DiffERR}.
\end{aligned}$$

That is, the difference between the estimated *load impacts* relative to two alternative baselines is the same as the negative of the difference between the *baseline errors*. This result points to the importance of baseline performance in calculating accurate load impacts.