The Wild West of Evaluation? Comparing Methods of Outlier Detection in Consumption Analyses

Jonathan Hoechst, Tetra Tech, Madison, WI

ABSTRACT

Identifying outliers, and the determination of which observations to include and exclude in a consumption data analysis is one of the most important decisions in any evaluation. Despite the direct influence these important choices have on results, this crucial step remains largely unstandardized across evaluation protocols and sectors.

A fundamental issue when conducting consumption analyses is dealing with accounts that exhibit extreme changes in energy use between the pre- and post-treatment periods. Generally, the approach is to remove accounts that experienced a change in consumption greater than a specified threshold, with the assumption that a factor outside the energy efficient intervention is driving the large change in energy use. Cases with excessively large changes in consumption, identified as outliers, can disproportionately influence and potentially bias results based on averages or regression analyses.

With consumption analyses involving only a small number of cases, individual cases with large changes in consumption can be examined, and a case-by-case determination can be made as to whether each record should remain in the analysis. When working with few cases, apparent outliers can individually be vetted to see if household characteristics, occupancy changes, or meter malfunctions explain their unusual consumption. This becomes impractical when the number of observations rises, leading to data screening rules and thresholds that generalize criteria for identifying outliers. While these rules can be effective at removing outliers, they are inherently subjective, and potentially determined arbitrarily. Researchers must determine the point at which changes in consumption become too extreme to reflect solely the installation of an energy efficiency measure(s).

This report consists of two sections. First, we conduct a review of studies to identify outlier detection methods used in the energy efficiency evaluation industry and other sectors, providing a summary of methods, thresholds, and underlying rationale supporting each method. We then apply the various outlier detection methods to a large set of residential advanced metering infrastructure (AMI) consumption data representing participants in energy efficiency programs to evaluate the effects of each outlier detection methods on results of the energy savings estimates.

Introduction

Background

As innovations like advanced metering infrastructure (AMI) become more widely adopted across the energy sector, industry professionals have more detailed consumption data than ever before about their customers. This wealth of information can present new insights into consumer behavior and make identifying and handling consumption outliers more crucial. These cases with extremely large changes in energy consumption between the pre- and post-treatment periods can potentially bias and muddle results, whether true outliers or an indication of technical malfunctions. Efficient strategies for identifying outliers can lead to more accurate results and a greater understanding of the impact and effectiveness of the energy efficiency measures being studied.

While efforts like the Uniform Methods Project for Determining Energy Efficiency Program Savings (UMP), CalTRACK, and state Technical Reference Manuals (TRM) aim to standardize energy efficiency evaluations, the subject of outliers is rarely explicitly addressed. The identification and removal of outliers have thus been inconsistent across organizations and evaluators. This can lead to discrepancies in savings calculations between utilities, implementers, and evaluators, especially in cases where the handling of outliers is not clearly documented in data cleaning methodologies.

Relatively little research has been done specifically addressing cases with extreme changes in energy consumption between the pre- and post-treatment periods in a billing analysis. Instead, discussion of outliers is confined to brief mentions during sections describing general data cleaning, and most analyses only address extreme measures of total energy consumption, not changes between treatment periods. In the following sections, we review and discuss some of the various approaches and methods used both in the energy industry and beyond.

Outlier detection and removal in consumption analyses are typically deployed by removing cases with extreme energy consumption, either during the pre-treatment period, post-treatment period, or both. This approach usually employs either:

- Numeric thresholds like minimum or maximum annual consumption limits
- Variance-based thresholds (removing those accounts that are more than three standard deviations from the mean annual consumption)
- Percentile thresholds (removing the top and bottom first percentile of consumption)
- A combination of these approaches

The rationale behind outlier detection and removal is that excessively low consumption could indicate that either a household is unoccupied, part of a multi-family building, or that meters are faulty, while those above the maximum may be extremely large homes or commercial properties.

A report prepared by the U.S. Department of Energy (DOE) utilized an approach combining a numeric threshold and a percentile threshold (DOE, 2014). In their example analysis, which evaluated the Better Buildings Neighborhood Program, the researchers first established a lower limit for plausible consumption indexed to 2.5% of the average annual single-family home energy consumption calculated using the Energy Information Agency's 2009 Residential Energy Consumption Survey (RECS). Next, they removed the accounts with total consumption below the first 0.5th percentile or above the 99.5th percentile. The researchers then reported the average electricity savings both with and without outliers included. In another report providing guidelines for energy efficiency analyses published by the British Department for Business, Energy & Industrial Strategy (Shah, 2021), a simple numeric threshold was used which retained any accounts with between 100 kWh and 25,000 kWh in annual consumption.

As this study is concerned with cases of extreme *changes* in consumption between the pre- and post-treatment periods, we will be focusing on studies that have addressed this specific issue. Researchers that have accounted for extreme changes in consumption follow a similar approach to those used for overall consumption. These primarily use either a numeric threshold, like removing any cases with energy changes over a certain limit, or a variance-based threshold that uses deviation from the mean or median changes in consumption. In most cases, outliers have been removed during data cleaning; however, in some cases outliers are not dealt with until the analysis phase when other factors that may influence extreme usage have been accounted for.

In a 2015 study of energy efficiency measures in the United Kingdom, researchers used a numeric threshold and set the maximum savings level at 50% (Adan and Fuerst 2015). The rationale provided was that such large changes were likely due to error rather than being attributable to the energy efficiency measures installed through the program.

An example study that controlled for consumption change outliers using variance-based metrics was published by the National Renewable Energy Lab (Belzer et al. 2007) that evaluated a joint initiative managed by the U.S. DOE and U.S. Environmental Protection Agency (EPA) that focused on improving whole-house energy efficiency for existing homes and examined roughly 7,500 homes managed by Austin Energy. In the study, the researchers employed a data cleaning metric that removed cases exhibiting post-treatment savings percentages that were more than three standard deviations from the mean. The researchers examined changes in weather-normalized consumption and provided regression results both with and without outliers removed. Another study published by the U.S. DOE in 2013 followed the same outlier identification and removal methodology (Hillman, 2013).

Looking outside of the energy domain yields a myriad of possible approaches to dealing with outliers; however, most are not relevant to the specific issue of extreme *changes* between pre- and post-periods. Many of those researching the subject of outliers continue to advance computational methods and approaches, notably using machine learning. A 2006 study proposed a method that used machine learning to generate scores indicating how much individual readings deviated from its normal consumption patterns (Seem 2006). The researcher compared meter readings in commercial buildings and identified extremely high and low consumption periods after accounting for time-of-day, weekends, and holidays. This allowed the researcher to identify and remove potential outliers based on the standardized score of how anomalous they were. Another 2015 study combined within-meter deviation with the behavior of nearby homes at the time of unusual readings to determine a joint “anomaly score” (Arjunan et al. 2015). They were then able to set thresholds for anomaly scores to identify and remove outliers. While these are beyond the scope of this paper as they do not pertain specifically to *changes* in consumption, they undoubtedly represent future directions for research in addressing outliers across fields and industries.

Methodology

Data Sources

This analysis includes data pertaining to a residential single-family home retrofit program implemented in the Southern United States between January 2017 and January 2020. The data used comes from the following four sources:

1. **Program Tracking Data:** We received program tracking data that contained account numbers, participation dates, addresses, measures received, reported TRM savings estimates for each measure received, and the utility associated with the account.
2. **Meter Data:** We received fifteen-minute interval data for residential customers from five utility companies for the period January 1, 2017 to January 1, 2020. This data contained an account number and kWh consumption value and timestamp (including the date, hour, and minute) for each fifteen-minute interval during the pre- and post-treatment periods.
3. **Weather Data**: This data was retrieved from the ASOS network and contained the hourly temperature readings for the period January 1, 2017 to January 1, 2020. We used data from the station closest geographically to each account, for a total of 59 weather stations.

Data Processing

The data cleaning steps outlined below were taken in the order written prior to the outlier identification process to ensure our data was of adequate quality for analysis. Accounts that met the following criteria were removed from the analysis:

1. **Solar participants**: accounts that have solar interconnection agreements. As these accounts produce some or all their own electricity, we would not have true consumption data.
2. **No meter readings**: accounts where meter data was missing entirely. It is not possible for us to include these accounts in the analysis.
3. **Insufficient metered period:** accounts where the earliest or latest meter reading date was less than 365 days from the participation date. In other words, accounts where the pre- or post-installation period was less than one full year are excluded. Using one full year of data both before and after project installation is standard practice and allows us to observe consumption in every season.
4. **Excessive missing meter readings:** accounts that were missing more than the equivalent of one total day of consumption data (missing more than 96 fifteen-minute meter data readings across the entire 730 days (365 pre and 365 post), not necessarily 96 consecutive fifteen-minute readings). This rule allows us to retain accounts with relatively small amounts of missing data, thus preserving the size and heterogeneity of the analysis group, while excluding those where large amounts of missing data could bias model coefficients.
5. **Excessive zero-kWh meter readings:** accounts with at least one week (672 fifteen-minute meter data readings) of continuous meter readings of zero kWh or at least one total month (2,880 fifteen-minute meter data readings) of meter readings of zero kWh, in aggregate. Long streaks or large amounts of meter readings of zero kWh indicate periods of vacancy, meter reading failure, or other issues that could bias model results. Meter readings of zero kWh are somewhat common (about 98% of accounts in the treatment group have at least one zero kWh reading); therefore, retaining accounts with some zero kWh readings was essential to preserve the size of the analysis group.

Table 1 provides a detailed breakdown of data attrition that occurred during each stage of the data cleaning process.

Table 1. Data attrition via data processing and cleaning steps

Step	Records remaining	Cumulative percent remaining
Census	33,567	100.0%
Solar	33,219	99.0%
No meter data	32,975	98.2%
Meter min/max < 1 year	32,963	98.2%
Missing Data	32,200	95.9%
0 kWh Data	28,783	85.7%
Final	**23,042**	**68.6%**

While the consumption data was measured in fifteen-minute increments, for purposes of identifying and removing outliers, usage data has been aggregated to 12-month periods immediately pre- and post-treatment. For analyses using linear regression, the original fifteen-minute consumption data was aggregated to the daily level.

Regression Models

We use a fixed-effect linear regression model to calculate savings estimates in both the original data and each of the outlier detection scenarios. We account for the effects of weather by calculating heating and cooling degree days. Heating degree days (HDD) are the difference between a reference temperature and the average daily temperature on a given day. The reference temperature represents the point at which heating equipment begins to operate. Cooling degree days (CDD) are the difference between the average daily temperature on a given day and a reference temperature that represents the point at which cooling equipment begins to operate.

$$HDD = ReferenceTemp - AverageDailyTemp$$

$$CDD = AverageDailyTemp - ReferenceTemp$$

The model for this analysis estimated average daily consumption and calculated the average daily HDD and CDD for each day for each account number. In this approach, each model allowed the heating reference temperature to range from 45°F to 65°F and cooling reference temperature to range from 65°F to 85°F for each household, in both the pre- and-post periods. The base temperature resulting in the best model fit (R^2) was assigned to the household, indicating which temperature reference point best explained the electricity usage patterns of each household. We then took the average household degree set-points for heating (56°F) and cooling (70°F) and used these to generate heating and cooling degree days across the population.

The regression model used the following specification:

$$Daily\ Consumption_{it} = \beta_1 HDD_{it} + \beta_2 CDD_{it} + \ \beta_3 post_{it} \ + \ \beta_4 HDD_{it} * post_{it} + \beta_5 CDD_{it} * post_{it} + esiid_{it}$$

Where for each customer' i' and day of the year' t':

$Daily\ Consumption_{it}$ = Actual daily consumption in the pre- or post- program period

$esiid_i$ = The participant account number, representing the daily kWh baseload for each account. Effectively, this is the intercept of account' i’

β_1 = The average change in daily usage resulting from an increase of one HDD in the pre-period

HDD_{it} = The base 56 heating degree days for the nearest weather station.

β_2 = The average change in daily usage resulting from an increase of one CDD in the pre-period

CDD_{it} = The base 70 cooling degree days for the nearest weather station.

β_3 = The average baseload savings in the post-period

$post_{it}$ = An indicator variable that equals 1 in the post-period (after the final measure installation for that account) and 0 in the pre-period (prior to any measure installation for that account

β_4 = The average savings in daily usage per HDD in the post-period

$HDD_{it} * post_{it}$ = An interaction term between HDD and the post indicator variable

β_5 = The average savings in daily usage per CDD in the post-period

$CDD_{it} * post_{it}$ = An interaction term between CDD and the post indicator variable

Once the model has been run for a data set, we fit the average annual TMY3 CDD and HDD to our model coefficients that contain the *post* term and multiply the *post* term by 365 since this coefficient is at the daily level. Summing those results yields our annual savings estimate.

Outlier Detection Techniques

Based on the existing literature, we chose three outlier detection techniques to test to measure the impact on savings calculations. Because the principal scope of this study is to identify and remove accounts with extreme changes in consumption, we chose not to focus on outliers of total consumption. We employed the following techniques: numeric thresholds, deviation-based thresholds, and percentile thresholds. For each technique, we applied three different threshold levels to test the impact on the regression analyses, as detailed below.

1. Numeric: accounts with changes over specific percentage thresholds (50%, 70%, 90%) are unlikely to be attributable to the measures installed and are likely the result of extraneous factors
2. Deviation-based: accounts with changes a certain number of deviations above or below the mean or median (1.5, 3)
3. Percentile: accounts with changes either in the top or bottom percentile ranges (0.25th, 0.5th, 1st)

Results

Summary Statistics

Table 2 provides an overview of the summary statistics from the original data set as well as those that result from each of the outlier detection methods applied (negative values indicate energy savings in the post-treatment period). Summary statistics for change in consumption by outlier detection method.

Compared to the data with all outliers retained, standard deviations were lower, and estimated savings were higher, likely due to the removal of few extremely high positive-change accounts that were identified by all our detection methods. Intuitively, many of the observations that are deemed outliers in one method are also identified by others. For example, an account that has a 75% decrease between the pre- and post-treatment periods is treated as an outlier by the 50% and 70% numeric threshold measures and may also fall within the 99th percentile for savings. Because many of the same observations are being removed, the means, medians, interquartile means, and standard deviations are all clustered relatively close together, indicating that the different approaches have similar effects on the data.

Table 2. Summary statistics for change in consumption by outlier detection method

Outlier Method	Records Retained	Records Removed	Percent Removed	Change in consumption (%)				
				First Quartile	Median	Third Quartile	Mean	Standard Deviation
None	23,042	-	0	-21.7	-9.3	2.6	-7.0	55.9
Over 50%	21,937	1,105	4.8	-21.0	-9.3	1.8	-9.0	18.2
Over 70%	22,690	352	1.5	-21.9	-9.5	2.0	-9.2	20.6
Over 90%	22,850	192	0.8	-21.9	-9.5	2.2	-8.9	21.7
Over 1.5 SD	22,786	256	1.1	-21.9	-9.5	2.1	-9.1	21.1
Over 3 SD	22,990	52	0.2	-21.8	-9.3	2.5	-8.1	23.8
Percentile 0.25th	22,926	116	0.5	-21.6	-9.3	2.5	-8.0	23.4
Percentile 0.5th	22,810	232	1.0	-21.6	-9.3	2.4	-8.2	22.2
Percentile 1st	22,580	462	2.0	-21.4	-9.3	2.2	-8.5	20.6

Figure 1 shows the distribution of consumption changes in the original data, as well as those retained under each of the different outlier regimes.

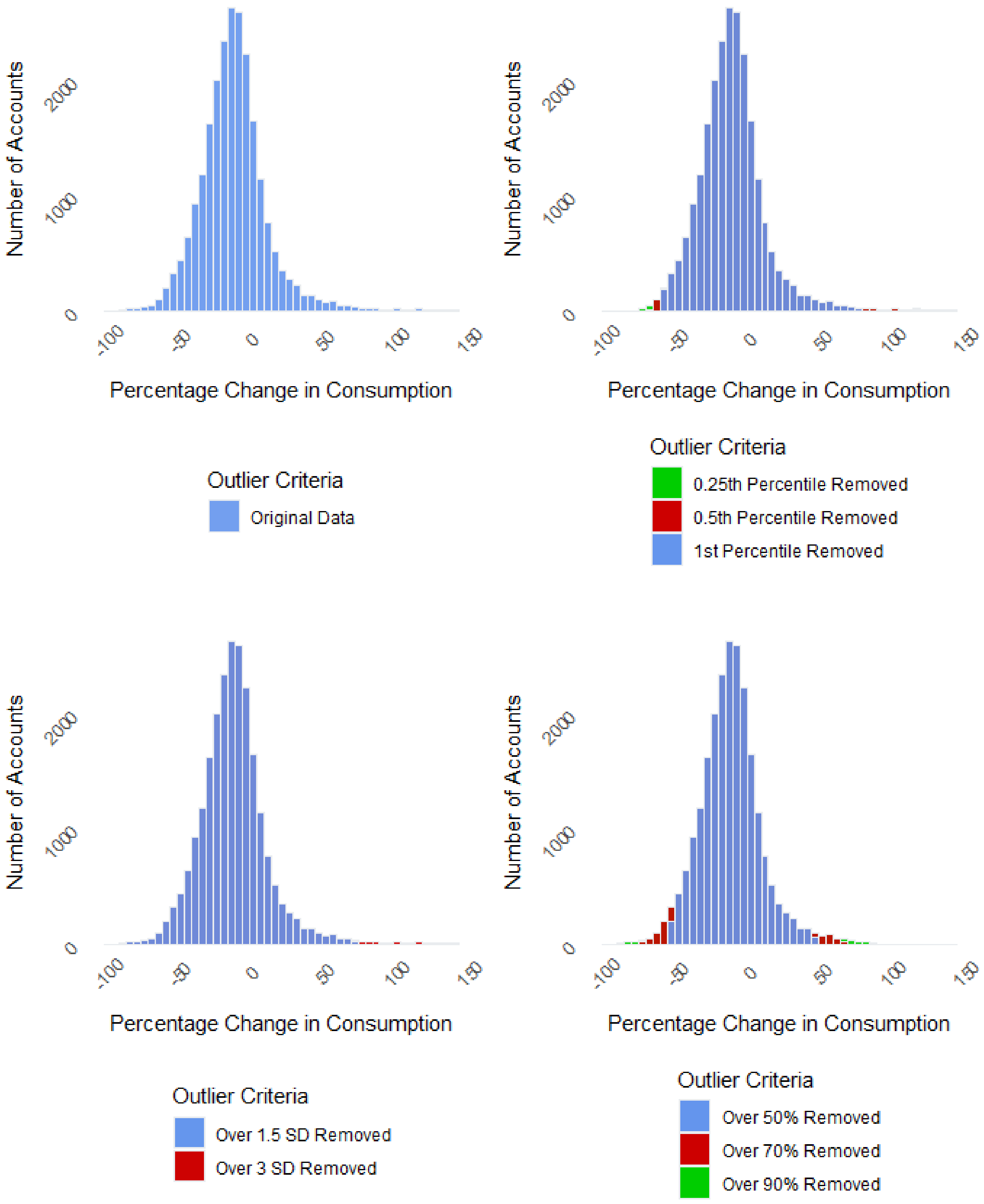

Figure 1. Percentage change from pre- to post-period across outlier regimes (from left to right, top to bottom: Normal, Percentile, Deviation, Numeric)

Regression Results

Modeled savings ranged from 7.3 percent in the original dataset to 7.9 percent of average pre-treatment period electric consumption. All model coefficients associated with the post-treatment period were statistically significant (at the 0.01 level). Table 3 presents regression results for the original data set as well as those that result from each of the outlier detection methods applied. It also provides aggregate consumption in the pre- and post-periods for each outlier method. Figure 2 provides a summary of total energy savings from the pre- to post-period using the different outlier methodology.

Table 3. Summary statistics for change in consumption by outlier detection method

Outlier Method	Average Pre-Period Consumption (kWh)	Total Pre-Period Consumption (kWh)	Total Post-Period Consumption (kWh)	Average Raw Savings as a percent of pre-period consumption	Average Modeled Savings as a percent of pre-period consumption*
None	15,145	348,978,471	311,873,350	9.2	7.3
Over 50%	15,356	336,856,502	300,799,761	8.9	7.7
Over 70%	15,267	346,399,637	307,608,258	9.1	7.9
Over 90%	15,223	347,841,439	309,155,737	8.1	7.8
Over 1.5 SD	15,246	347,389,981	308,331,930	8.0	7.9
Over 3 SD	15,170	348,766,001	311,072,491	8.2	7.5
Percentile 0.25th	15,178	347,977,158	310,801,686	8.5	7.4
Percentile 0.5th	15,204	346,803,543	309,761,329	9.2	7.4
Percentile 1st	15,249	344,314,722	307,508,530	8.9	7.4
*All regression coefficients associated with the post-period were statistically significant					

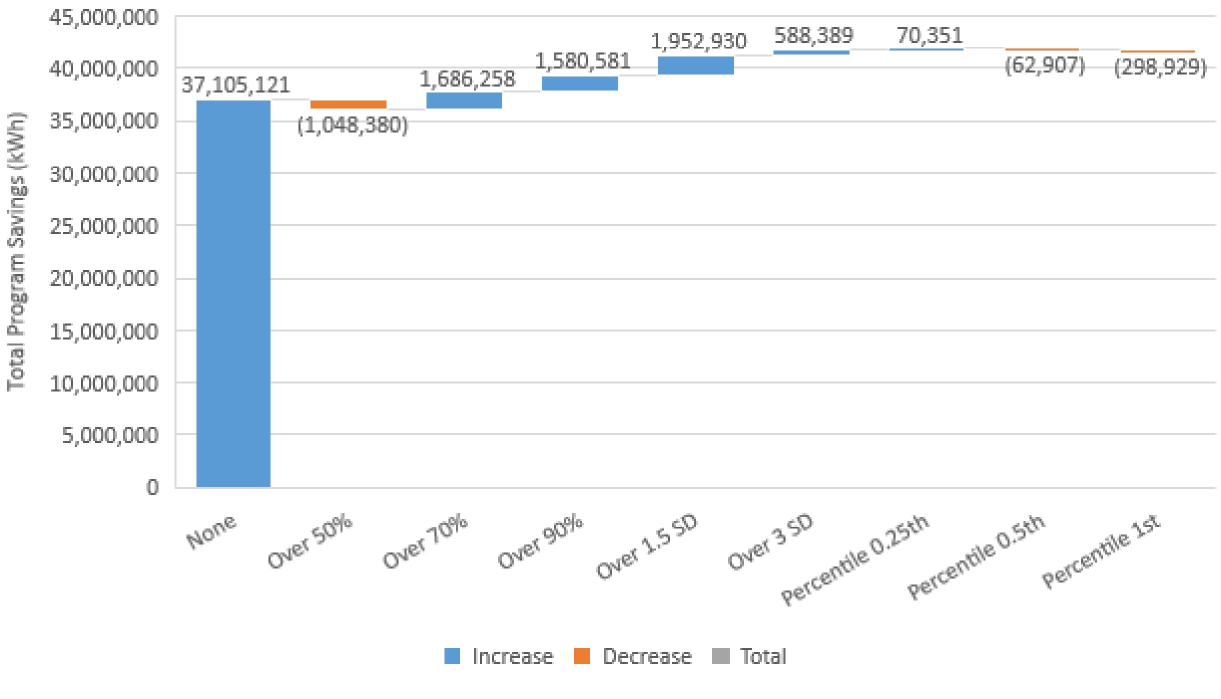

Figure 2. Total energy savings from pre- to post-period across outlier methods

Conclusion

The results of this study indicate that the various approaches tested did not drastically change the character of the savings calculations. It is notable that all our outlier regimes lead to different estimated savings, although these ranged only between 0.1 and 0.6 percent of pre-period consumption. We note that the increase in savings that results from our outlier detection methods likely stems from unique features of this data set that are likely not generalizable to other efficiency analyses. The specific consumption data used contains more cases of accounts with extreme increases in consumption compared to extreme decreases in consumption, which may not be true as a rule in other studies.

A limitation of this study is that we lack objective metrics for comparing the effectiveness of each outlier detection method. While we can compare the effects of different methods, we cannot determine which is best to use in which situations. As we described earlier in this paper, the energy industry lacks widely recognized criteria for evaluating outlier detection methods, hence the varied approach across studies. This inherently presents challenges in recommending one approach over the other. We also did not examine the effects of combining methods, as is sometimes done in other studies and may be appropriate depending on the situation.

Another limitation of this study is that its results pertain primarily to large-scale consumption analyses that are based on regressions or average changes in coefficients. In these types of cases, the assumption can be made that it is safer to err on the side of removing more potential outliers as home retrofits are unlikely to be the root cause of extremely large shifts in consumption from one year to the next. There are many cases, especially those related to programs with performance incentives, where it is very important to retain and study what would normally be considered outliers.

One potential direction for future research may be to use changes in weather-normalized consumption to initially determine outliers rather than raw data. This would allow researchers to control for weather-related consumption shifts (for example, the post-period has significantly higher temperatures on average during the summer compared to the pre-period) that could cause accounts to be misidentified as outliers. A drawback to this approach is that it may require greater computational resources to weather-normalize a larger population, some of which will be removed from the final analysis.

Another approach that could be studied is to take advantage of the AMI data available and use machine-learning to determine outlier observations within individual accounts' consumption. This could involve training an algorithm to learn individual account consumption patterns and identify specific anomalous observations that deviate from these patterns. This could be particularly useful for identifying malfunctioning meters or appliances/electronics that are behaving improperly which may have anomalous readings that are smoothed by only examining aggregating consumption. It could also identify and remove individual outlier meter readings so that more accounts could be retained for the final analysis. Researchers may consider applying this approach to a smaller data set initially to test its performance, though, as it would be much more computationally intensive than studying annualized consumption.

References

Adan, H. and F. Fuerst. 2015. "Do energy efficiency measures really reduce household energy consumption? A difference-in-difference analysis." *Energy Efficiency* 5: 1207-1209. https://doi.org/10.1007/s12053-015-9418-3.

Arjunan, P., H. Khadilkar, T. Ganu, Z. Charbiwala, A. Singh, P. Singh. 2015. "Multi-User Energy Consumption Monitoring and Anomaly Detection with Partial Context Information." *BuildSys '15: Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments:* 35-44. https://doi.org/10.1145/2821650.2821662.

Belzer, D., G. Mosey, P. Plympton, and L. Dagher. 2007. "Home Performance with ENERGY STAR: Utility Bill Analysis on Homes Participating in Austin Energy's Program." National Renewable Energy Laboratory (NREL). https://www.nrel.gov/docs/fy07osti/41903.pdf.

DOE (U.S. Department of Energy). 2014. "Guide for Benchmarking Residential Energy Efficiency Program Progress." https://www.energy.gov/sites/default/files/2014/11/f19/bbr_program_benchmarking_guide_draft_nov2014_0.pdf

Hillman, T. 2013. "Small Town Energy Program (STEP) Assessment of Program Impacts through Utility Bill Analysis." U.S. Department of Energy. https://www.energy.gov/sites/prod/files/2014/09/f18/G5d%20STEP%20Utility%20Bill%20Analysis%20Report.pdf.

Seem, J. 2006. "Using intelligent data analysis to detect abnormal energy consumption in buildings." *Energy and Buildings* 39: 52-58. https://www.researchgate.net/publication/245196873_Using_intelligent_data_analysis_to_detect_abnormal_energy_consumption_in_buildings

Shah, H. 2021. "National Energy Efficiency Data-Framework (NEED): Summary of Analysis, 2021." Department for Business, Energy & Industrial Strategy. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1008681/need-report-2021.pdf.