RECURVE
SHAPE THE FUTURE OF ENERGY

GRIDmeter 2.0: Comparison Groups for the COVID Era and Beyond

Updated: 3/18/2024

Acknowledgments

This report was developed based upon funding from the Alliance for Sustainable Energy, LLC, Managing, and Operating Contractor for the National Renewable Energy Laboratory for the U.S. Department of Energy.

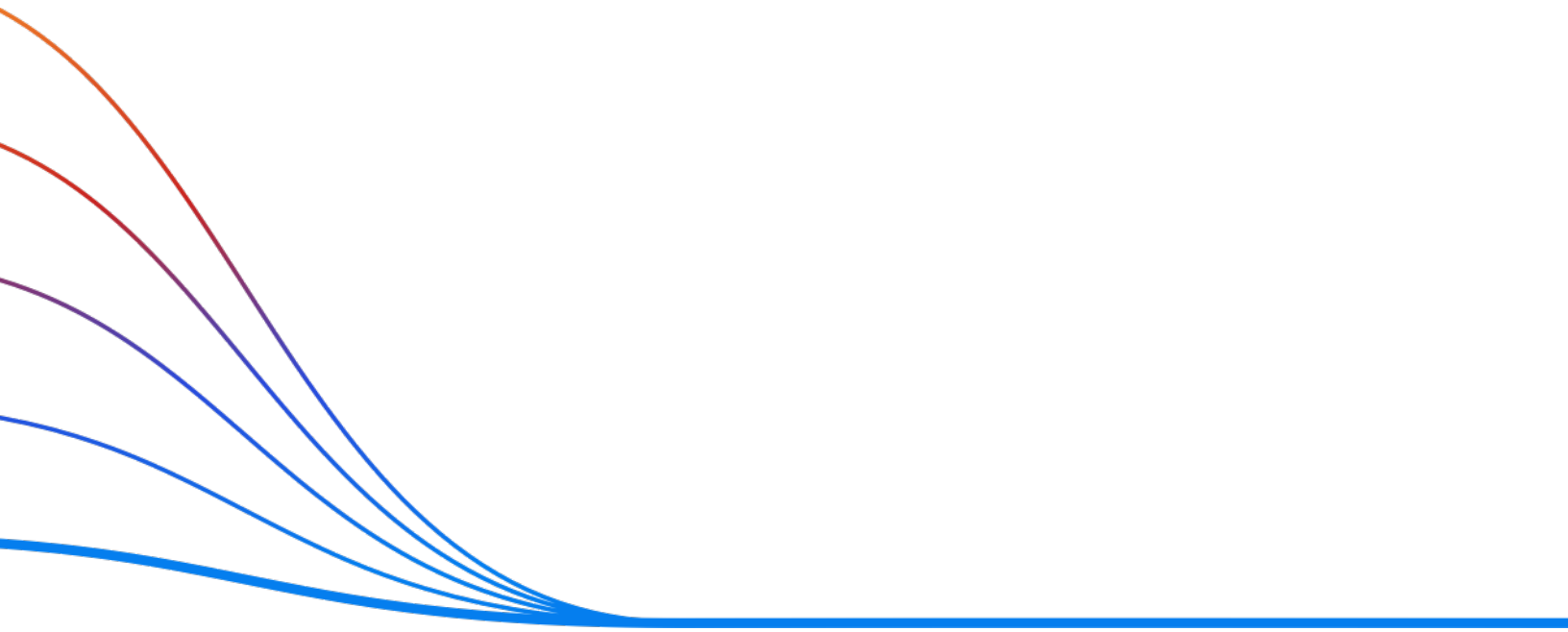
Recurve would also like to thank MCE for supporting this project by providing secure access to data. Without secure data-sharing partnerships between utilities/energy providers and the demand side service industry, the next generation of programs capable of fighting climate change, enhancing grid resilience, keeping rates affordable, and meeting customer needs simply cannot be developed. MCE's partnership in this effort shows it is serious about solving these issues and helping others do their part.

Finally, Recurve thanks the members of the Comparison Groups Working Group, who devoted their time and effort to listening, reviewing, and providing feedback throughout the research and development of the methods and recommendations in this report. Through nearly a dozen working group meetings and outside engagement, the working group members helped focus our efforts and ensure a final product that we believe can genuinely help the industry as we continue to address COVID and seek to modernize demand-side programs.

RECURVE

SHAPE THE FUTURE OF ENERGY

Chapter 1: Standardizing Comparison Groups to Enable Demand-Side Programs to Compete at Scale



Background

This document describes methods for 1) selecting a comparison group of meters belonging to buildings that are not currently participating in demand-side energy programs; 2) calculating the change in energy consumption in a way that allows the effects of exogenous factors to be accounted for in buildings that are participating in demand-side energy programs; 3) using the results of comparison group savings calculations to correct for model error and adjust the savings of program participants.

Common comparison group methodologies are described in the Uniform Methods Project, Chapter 8, "Whole Building Retrofit with Consumption Data Analysis Evaluation Protocol."[1] Program evaluations rely on comparison groups to adjust savings calculations to account for non-program effects on energy consumption. These effects can include program "free-ridership" wherein program participants leverage rebates for projects they would have undertaken in the absence of a program. In contrast, programs that pay for savings based on the metered performance of portfolios of buildings require a more focused calculation, namely one that utilizes a comparison group to adjust for intrinsic model error as well as exogenous effects on building energy consumption.

In the guidance that follows, many of the concepts of comparison group methodologies will seem familiar. The 'in-flight' comparison groups described here are designed to support population-level programs measured at the meter in the normal course of program operation. Many emerging programs also utilize pay for performance structures in which whole-building meter-based savings calculations aggregated across a portfolio of projects inform an aggregator payment directly. As such, both the requirements and the embedded assumptions about the purposes of a comparison group will differ from what is found in UMP Chapter 8 and other similar program evaluation protocols.

In-flight comparison groups are distinct from evaluation-based comparison groups in two key respects. First, the initial construction of an in-flight comparison group is inherently naive to the construction of the treatment group. Unlike opt-out programs that enroll customers all at once and for which a comparison group can be selected in advance of program enrollment based on selected participants, opt-in programs that enroll customers throughout a program term will not have a complete accounting of participants until after enrollment has closed. Second, evaluation-based comparison groups often attempt to match non-participants based on a variety of similarity functions, including socio-demographic characteristics. The data collection costs of this practice limit the frequency of this type of evaluation and render it both impractical and infeasible for rapid deployment during program operation. As a result, the conclusions that can be drawn from in-flight comparison groups may be more limited than what might be derived during a more comprehensive impact evaluation conducted at a later stage.

The primary purpose of an in-flight comparison group is to account for the effects of systemic changes in energy usage unrelated to program participation. Examples of this type of systemic change in consumption would include reduced usage related to fuel shortages (rationing), reduced usage related

[1] https://www.nrel.gov/docs/fy18osti/70472.pdf

to rate changes, as well as the obvious and motivating reason for the development of these methods, which is the change in consumption patterns in response to the emergence of COVID-19.

For an individual building, there is no such thing as a pure exogenous effect. There is only the way in which exogenous factors interact with the drivers of energy consumption within that building. Just as there is some degree of uncertainty with respect to the causality of energy savings within a building in the first place, there will be an accompanying degree of uncertainty with respect to the effects of exogenous factors on the change in energy consumption within a building. Two buildings installing the same measures could see different savings under normal conditions and even greater differences under the strain of COVID-19. It is both impractical and infeasible to try to disentangle the unique ways in which exogenous factors interact with energy use patterns at a building level. Instead, the larger purpose of enabling scalable demand-side programs must be to capture exogenous effects at a portfolio level. Individual differences can fade to reveal a broader trend amongst a treated set of customers.

The methods described below are intended to enhance OpenEEmeter[2] methods for calculating whole building energy savings. Unless otherwise noted, assumptions about baseline conditions, modeling, data requirements, and more are based on the expectation that avoided energy use will be calculated at the site level following OpenEEmeter specifications. Alternate approaches to calculating site-level or aggregated savings may contain implicit assumptions that negate the value of the comparison group methods in this guidance.

This guidance has not attempted to reconcile the avoided energy use calculation that relies on the actual weather of the reporting period with evaluation approaches that calculate energy savings under the conditions of a "typical weather year." There are challenges associated with COVID-related changes in energy consumption that complicate efforts to "normalize" savings to a typical year (whether normalizing weather or consumption). This topic will require additional research and methodological guidance beyond the scope of this project.

***

A comparison group should have a primary objective: identify a set of non-participating buildings likely to exhibit similar model error trends (on account of intrinsic and exogenous factors) as the program participants. However, this selection process is challenging for three reasons. First, different types of exogenous factors can lead to different types of responses. For example, a service territory-wide switch to a time-of-use rate structure would be expected to impact the energy usage patterns across the entire population. However, one might expect that income-sensitive customers with high peak-load consumption would be more sensitive to a time-of-use rate than a typical customer. Similarly, COVID-related impacts might be felt most acutely amongst customers, both residential and commercial, with greater work-from-home flexibility.

A second challenge associated with comparison group selection is that while comparison groups can be constructed based on historical data, exogenous events might introduce a new divergence between

[2] https://github.com/openeemeter/eemeter

treated and comparison group buildings. For example, COVID-related energy changes due to business shutdowns were more extreme in certain small business sectors than in certain “essential” businesses, despite a broad similarity in consumption patterns prior to COVID that would otherwise indicate a good comparison group match.

A third challenge is the limited availability of data that might help account for differing exogenous effects. While, with the right data, we might be able to perform some filtering, such as classifying buildings according to their business type, other filters such as trying to determine which residential homes are adding occupants and which are losing occupants, for example, would be much more difficult to construct.

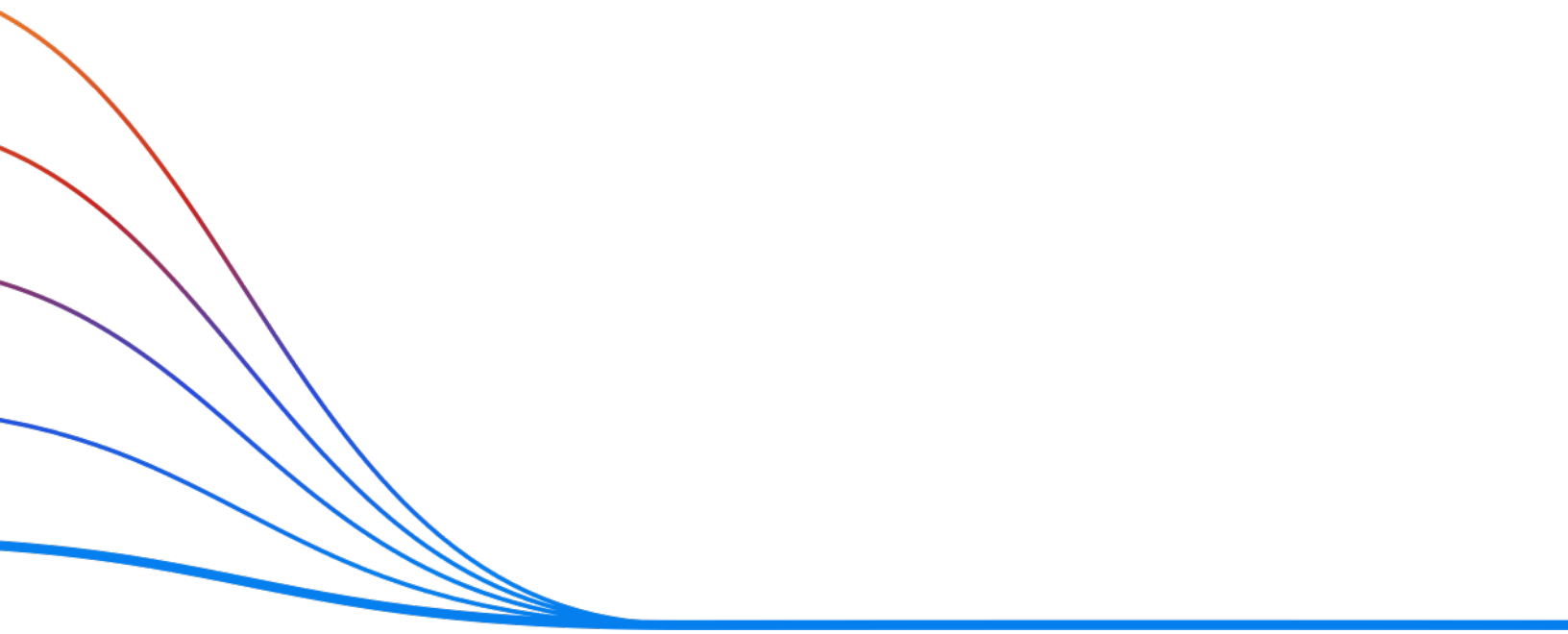
Along with the in-flight nature of a comparison group needed to facilitate meter-based programs, these three factors - dissimilar responses to exogenous factors, unpredictable exogenous events, and limited data for assigning buildings to cohorts - require the comparison group selection process to utilize a more standardized and consistent methodology than what might be found in traditional impact evaluations.

The methodological approach outlined here prioritizes replicability and universality, recognizing that under certain circumstances there will be a preference for waiting until after program participants have enrolled or for finding additional data about participants and non-participants to account for differential responses to exogenous conditions.

Many of the methods detailed below stem from the results of analysis conducted with the support of MCE. Without MCE’s support to provide data for this project this effort would not be possible, and we thank MCE for helping the entire demand side industry take on one of the more unique challenges in recent times.

RECURVE
SHAPE THE FUTURE OF ENERGY

Chapter 2: Key Analyses and Results That Inform Methods Development



I. Dataset

The dataset utilized for this study is composed of random samples of 100,000 residential and 22,407 non-residential non-participant non-solar meters across MCE's service territory. All meters had at least 85% complete hourly electric traces from March 1, 2019 through August 22, 2020. In addition, metadata was provided that allowed Recurve to map solar PV status, climate zone, business type, and a variety of additional site-level information. MCE territory includes California climate zones 2, 3, and 12.

II. COVID Impacts

As a basis for approaching comparison groups in the era of COVID, we must first understand the size, scope, and variability of these impacts and how they differ between customer segments. Therefore, as a first step of this effort we have measured the change in electricity consumption attributable to COVID for all meters in MCE's service territory. To make this measurement we performed OpenEEmeter 2.0 hourly calculations for each meter using the following baseline and reporting period timeline:

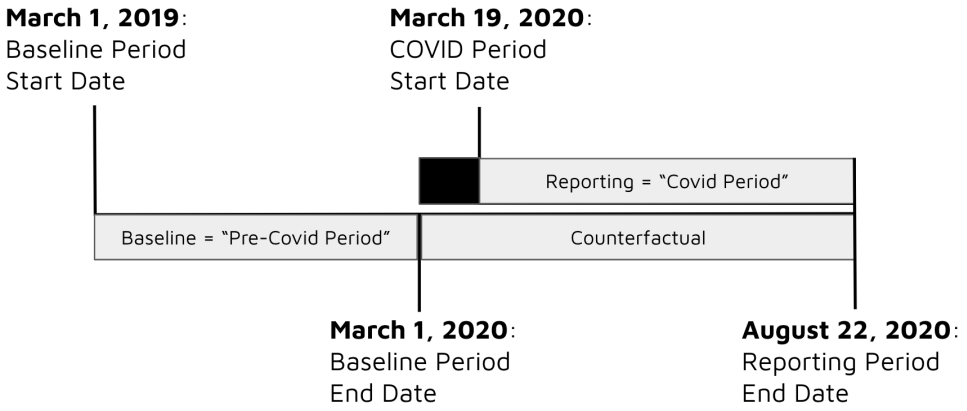


Figure 1: Metering timeline for the calculation of COVID impacts. This timeline is used for many of the comparison group tests described throughout this document.

March 19 is chosen as the COVID-period "start date" because on that date California entered a state-mandated stay-at-home order, which has remained in place to varying degrees to the time of this writing. The baseline period has been chosen as the 366 days leading to March 1, 2020. With this timeline, the COVID shutdown is essentially treated as though it were a program intervention in a typical meter-based savings calculation for a demand-side program. The baseline period model, developed from a year of "pre-COVID" data, is projected forward as the counterfactual into the COVID period and associated impacts are determined for each meter by comparing observed usage to the counterfactual predicted usage. The OpenEEmeter methods account for temperature and these calculations are thus weather-normalized.

We note that in both existing and future programs, the impacts of COVID may be entirely or predominantly in the baseline period, reporting period, or could be in both. While it is not feasible to extensively test each of these scenarios, the metering timeline of Figure 1 provides a clean view into a relevant yet limiting case in which the entirety of the baseline period is not affected by COVID while the reporting period contains what is anticipated to contain the most severe COVID impacts.

In order to enable clear outcomes, we will focus only on non-solar meters for all analyses throughout this work.

A. Residential Sector

Looking first at the Residential sector, Figure 2 shows the observed and counterfactual daily load shape for an average meter.

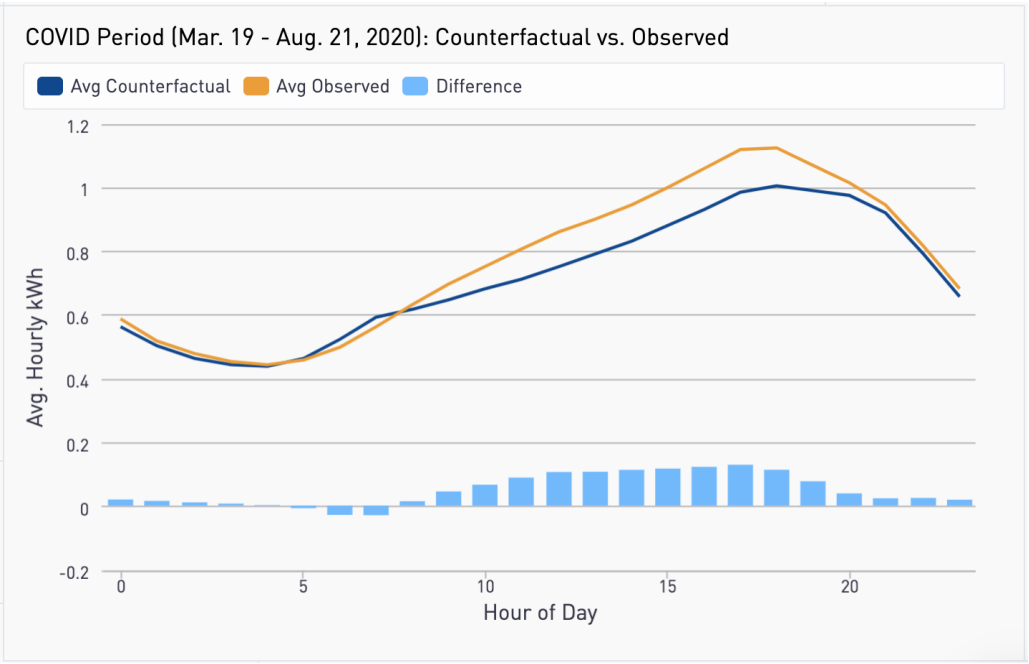


Figure 2: Average observed and counterfactual daily load shapes for MCE non-solar Residential customers during the COVID period.

We measure a total increase in consumption of 7.9% due to COVID, with most of this increase occurring in the mid-day hours. These results are intuitive given that many customers who would have been away at work have needed to stay home.

While the average customer experienced an increase in usage, we observe a wide distribution in the COVID impact measurement at an individual customer level. Figure 3 shows the distribution of COVID impacts across the residential sector.

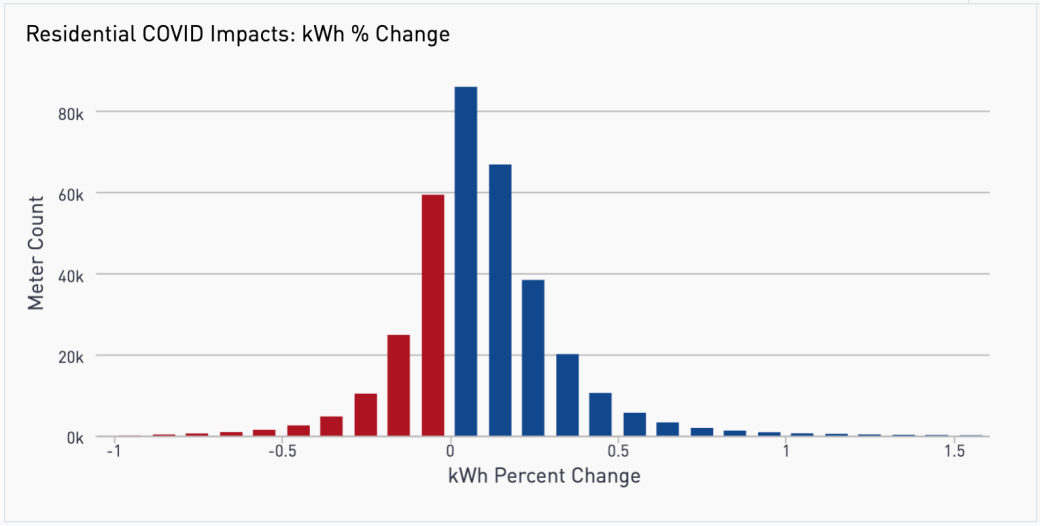


Figure 3. Distribution of the percent change in electricity consumption due to COVID for MCE non-solar Residential customers.

Despite the wide distribution among individual Residential customers, we have observed relatively little change in this distribution among different demographic segments of the population, including when isolating particular geographic locations and assessing the low-income sector. In addition, the distribution of Figure 3 is largely stable against different usage characteristics that we have tested. More detailed COVID impacts results for the Residential sector can be found in Chapter 5 and Appendix B.

B. Commercial Sector

Figure 4 shows the average observed and counterfactual daily load shapes for non-solar Commercial customers in MCE territory. A 15% overall reduction in electricity consumption is observed.

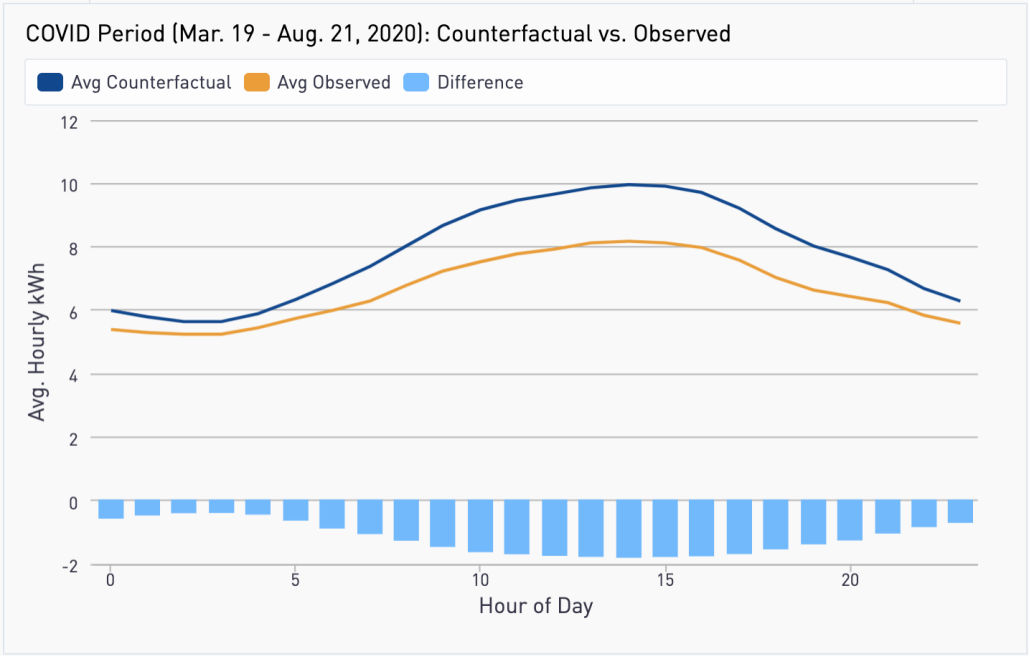


Figure 4: Average observed and counterfactual daily load shapes for MCE non-solar Commercial customers during the COVID period.

As in the Residential sector, the largest difference is seen in the middle of the day where most businesses have typical operating hours.

In assessing COVID impacts in the Residential and Commercial sectors, we observe an additional commonality and one important difference that has implications for comparison groups:

- As with Residential, a wide distribution of COVID impacts exists at an individual customer level among different businesses.
- Unlike Residential, we observe that different segments of the Commercial sector exhibit widely different responses to COVID.

As an example, Figure 5 shows distributions of COVID impacts for Grocery and Convenience stores (left) and Hotels and Lodging facilities (right).[3]

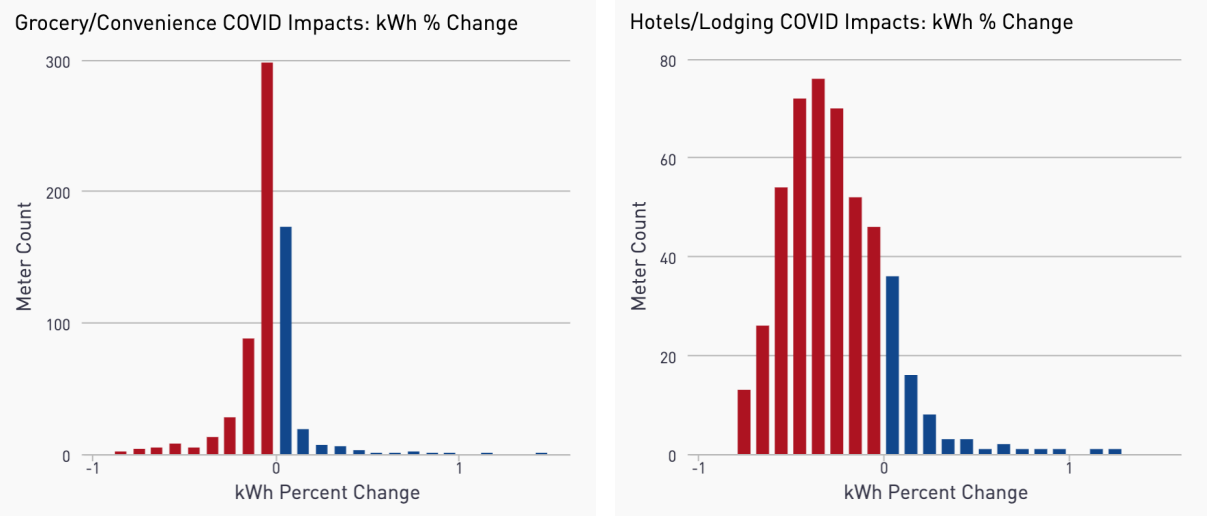


Figure 5: Distribution of the percent change in electricity consumption due to COVID for Grocery and Convenience stores (left) and Hotels and Lodging facilities (right).

While the Grocery/Convenience stores have seen an 8% decrease in consumption, the business operations of Hotels have been far more impacted with a 24% drop in electricity usage. While these are just two examples, across the distinct economic segments of the Commercial sector we observe a wide range of COVID impacts (full results in Table 1 below).

If creating a comparison group that is blind to business type, savings calculations are likely to be subject to significant error on account of differing responses to COVID. Figure 6 shows how this effect plays out for the same segments: Grocery/Convenience (top) and Hotels/Lodging (bottom). This figure shows the

[3] A key tool for this research is the categorization of MCE Commercial costumes into "NAICS Groups," which yield high-level business type assignment. Appendix 1 provides more detail on the mapping procedure to establish these NAICS Groups.

results of difference of differences calculations when taking samples from these subsectors as a "treatment" group and utilizing a random sample of commercial customers as a comparison group. The vertical dotted line indicates March 19, the start of the COVID period. At this point, the random sample does not effectively mirror the response to COVID unique to these business types and the effects are observed as residuals that are consistently low (Grocery/Convenience) or high (Hotels/Lodging).

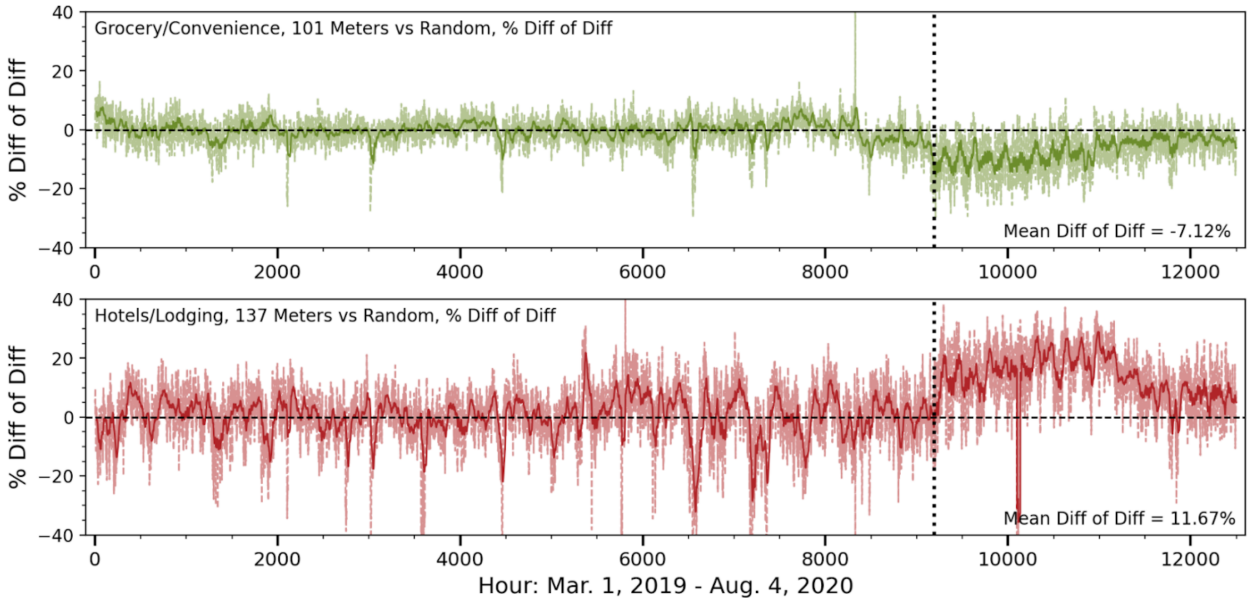


Figure 6: Difference of differences calculations throughout the baseline and COVID periods for the Grocery/Convenience segment vs. a random sample of commercial customers (top) and the Hotels/Lodging sector vs. a random sample of commercial meters (bottom). The dotted line indicates the start of the COVID period.

As detailed in Chapter 6/Appendix C, with data on business type, comparison groups can be formed that can neutralize the large between-segment differences observed in the Commercial sector. However, where business type information is not available other strategies must be employed, either on the measurement or program side, to ensure reliable measurement can be made to appropriately account for COVID impacts within meter-based Commercial programs.

One such M&V approach would be to identify usage characteristics observable in a baseline period that are predictive of customer responses to COVID. For instance, it may be that customers with higher total baseline period usage tend to be impacted less by COVID. However, as detailed in Table 6, we have been unable at this point to find consumption characteristics that are adequately predictive of COVID impacts to eliminate the business type differences we observe.

III. Comparison Group Sample Size

A foundational element of comparison group selection is the proper sizing of the sample. If a sample is too small, it will introduce undue uncertainty into the savings calculation simply by random noise effects. However, if a sample is larger than needed, program administrators may release more non-participant records than justified, the comparison group may have to be made less representative of a treatment group, and computational costs will be higher than necessary.

To gauge the degree to which random variability in comparison group selection can produce uncertainty in the calculation of savings, we performed the following analysis for the Residential sector:

1. Compiled two non-overlapping random samples of 50,000 MCE residential meters
2. Performed OpenEEmeter 2.0 Hourly calculations using the metering timeline of Figure 1 for each meter.
3. The first of the random samples is taken as the “treatment” group. The second random sample is taken as a comparison pool.
4. From the comparison pool 50 random samples each are pulled for sample sizes of 100, 250, 500, 1000, 3000, and 10000 meters, respectively.
5. Differences between the treatment group and comparison samples are calculated on the bases of pre-COVID observed kWh, pre-COVID model kWh, COVID period observed kWh, and COVID period counterfactual.

Because both the treatment group and the comparison pool are selected at random from MCE’s customer base, the expected value of these differences is 0 and residuals are attributable to random variation. Figure 7 shows the results of this analysis for a baseline OpenEEmeter model.

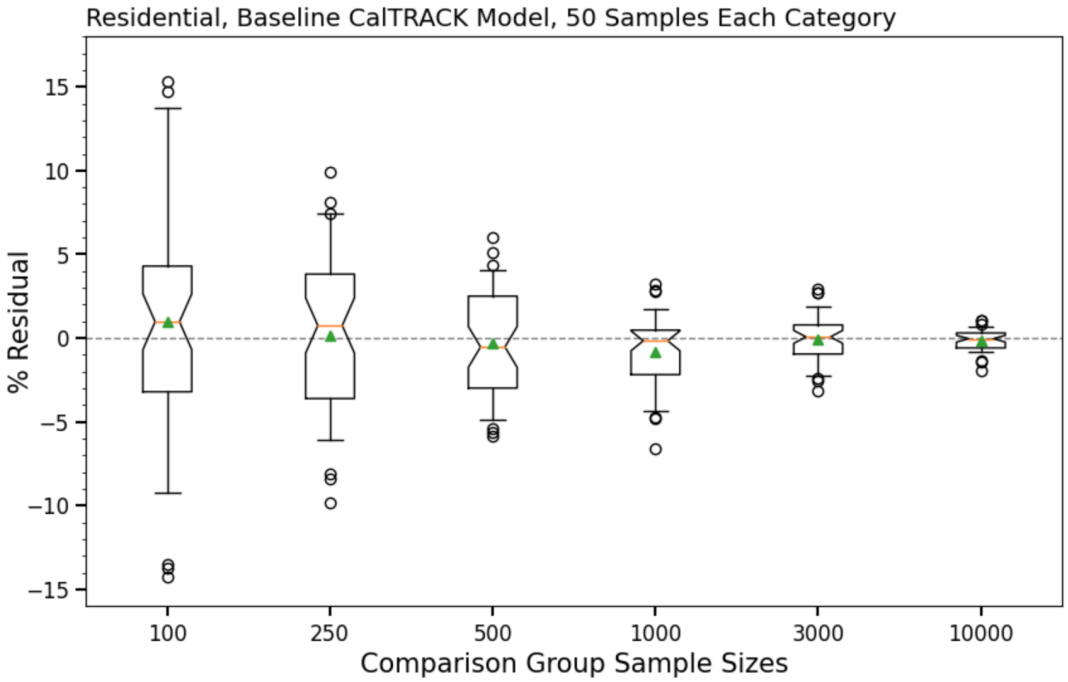


Figure 7: Box and whisker plot showing the distribution of residual error introduced by different comparison group sample sizes for Residential customers in MCE’s service territory. Each box represents the interquartile range and the whiskers represent the 0.05 - 0.95 probability range. Outliers are shown as individual data points outside the whiskers. The mean for each sample size is shown as a green triangle, the median an orange bar, and the notch is set to show the 95% confidence interval of the mean.

The results of Figure 7 imply that comparison group sample sizes of 500 or less in the residential sector are prone to introducing uncertainty on the order of 5% or more. Sample sizes of 3,000 or more can

reduce uncertainty to +/- 2% in the vast majority of cases and yet larger samples reduce variance to an even greater degree.

As the next step in this analysis we investigated each element that feeds into a difference of differences (savings) calculation.[4] Results are given in Figure 8 for the 3,000-meter sample size.

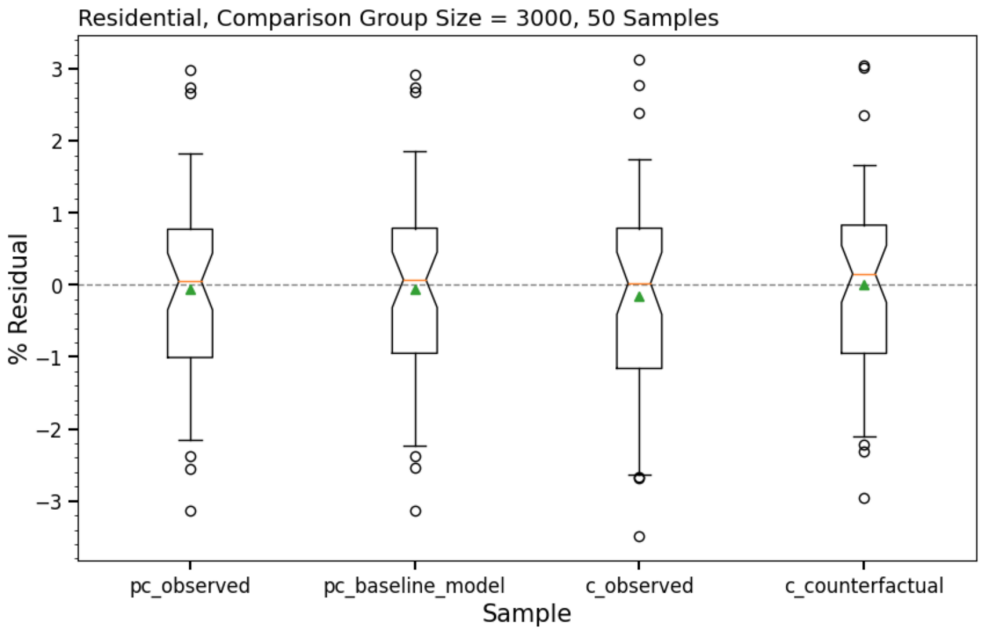


Figure 8: Box and whisker plot showing the variability present across each component of the difference of differences calculation for the 3000 sample size. “pc” indicates pre-COVID and “c” indicates COVID.

Figure 8 shows a high degree of consistency between the various elements of the difference of differences calculation. In both pre-COVID and COVID periods, nearly the same degree of variability is apparent between observed and modeled components. This is the case for all sample sizes investigated.

Given these results, we recommend a comparison group size of at least 3,000 meters for most Residential programs and for Commercial programs if possible. Fewer meters may be appropriate for deep retrofit programs expected to deliver more than 15% savings, while programs expecting under 5% savings may require larger groups. As noted above, larger comparison groups generally necessitate larger comparison pools to produce the same quality match to a given treatment group.

For Commercial sector programs the number of comparison group meters needed for an appropriate level of uncertainty will depend on the targeted segment(s). Some business types tend to have far more stable, consistent, and predictable usage patterns while others have a higher degree of diversity and variation in energy usage between customers, which leads to the need for both larger treatment and comparison groups to achieve the desired level of measurement precision. While it is beyond the scope of this work to conduct a detailed assessment of variability within every commercial subsector, we have conducted tests that can lend a foothold to the question of sample size.

[4] Details of the difference of differences calculation are given in Chapter 4.

Table 1 gives results for tests in which the available meters within each commercial segment are split evenly with the resulting samples compared against one another. For each NAICS group the sample sizes are listed. The % Diff columns show the percentage difference for each sample between the OpenEEmeter prediction and the observed value for total consumption in both the pre-COVID (baseline) and COVID (counterfactual) periods. The discrepancy observed between samples for each subsector is shown in the "Sample 1 - Sample 2" columns.

Table 1

			Pre COVID		COVID	
NAICS Group	Sample Size	Sample	% Diff	Sample 1 - Sample 2	% Diff	Sample 1 - Sample 2
Administrative/Civil	1190	1	-0.11	-0.04	-9.21	2.39
		2	-0.07		-11.60	
Automotive	439	1	-0.05	-0.06	-8.63	-2.23
		2	0.01		-6.40	
Banks	94	1	-0.12	-0.03	-7.18	-0.30
		2	-0.09		-6.88	
Beauty	476	1	-0.07	-0.01	-60.42	-1.64
		2	-0.06		-58.78	
Churches/Religious	283	1	-0.26	-0.18	-29.64	2.22
		2	-0.08		-31.86	
Construction/Contractors	470	1	-0.04	-0.01	-10.58	-1.04
		2	-0.03		-9.54	
Fitness	140	1	0.14	1.06	-49.78	2.68
		2	-0.92		-52.46	
Grocery/Convenience	101	1	0.15	0.15	-8.19	-1.49
		2	0.00		-6.70	
Hotels/Lodging	137	1	-0.22	-0.10	-26.97	-5.56
		2	-0.12		-21.41	
Medical_Offices	525	1	-0.01	0.00	-19.19	-3.79
		2	-0.01		-15.40	
Offices	554	1	-0.03	0.04	-21.02	-3.07
		2	-0.07		-17.95	
Real_Estate	1208	1	0.03	-0.02	-15.17	-0.05
		2	0.05		-15.12	
Restaurants/Bars	436	1	-0.01	-0.09	-22.16	-2.69
		2	0.08		-19.47	
Retail	663	1	-0.08	-0.07	-20.35	2.12
		2	-0.01		-22.47	
Schools	53	1	0.11	0.11	-44.88	-4.64
		2	0.00		-40.24	
Unassigned	5356	1	-0.09	-0.06	-11.21	-0.57
		2	-0.03		-10.64	
Warehousing/Postal	61	1	0.11	-0.01	29.57	27.09
		2	0.12		2.48	

All subsectors show minimal divergence in the baseline period. Most subsectors exhibit a discrepancy of under 3% during the COVID period. For 16 of the 17 subsectors this value is under 6%. While some of this could be luck of the draw as we are only taking one arrangement of each sampling split, the low

discrepancies, even with most sample sizes well under 1,000 meters, indicate that with business type information reliable comparison groups can be formulated for the commercial sector.

Instead of issuing a formal recommendation on sample size for all subsegments of the Commercial sector we provide the following consideration: Most jurisdictions have relatively few Commercial meters relative to Residential accounts, and the number of meters available for any particular economic segment is likely to be very limited. To facilitate reliable comparison group formulation, utilizing all non-participating meters that correspond to a program's participant group would provide the most statistical power possible for measurement during the COVID era. Clearly, any customer pulled into the program should be tracked accordingly and removed from the comparison group. If more meters are available or if refined sampling is still desired, additional sampling can be done as described in Chapter 3.

IV. Hourly Measurements

The reliable measurement of load impacts on an hourly basis is critical for many modern demand flexibility programs. In addition, programs are becoming more targeted, where customers with particular usage characteristics, like high cooling loads or peaking load profiles, offer an opportunity to enhance the cost-effectiveness and scalability of demand-side programs. With these trends in mind, we have assessed the sensitivity of hourly measurements to a comparison group.

Figure 9 gives an example of an hourly difference of differences calculation in which the "treatment" group consists of 3,000 meters that exhibit a shallow evening ramp. The top plot shows results when this sample is tested against a random sample of residential customers. The bottom plot gives results when the comparison group is instead pulled from customers with similar load profiles.

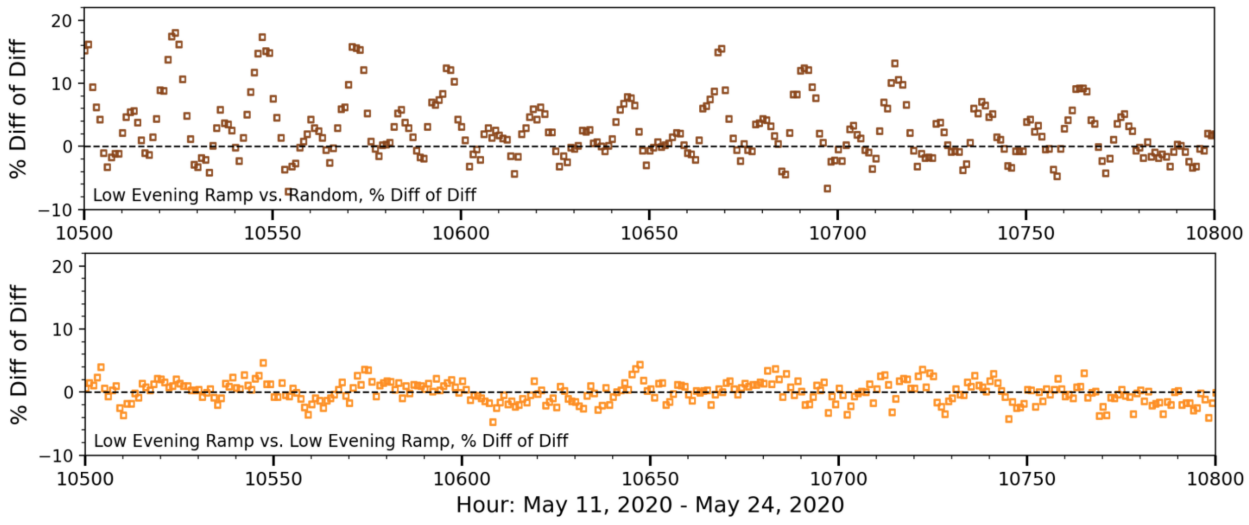


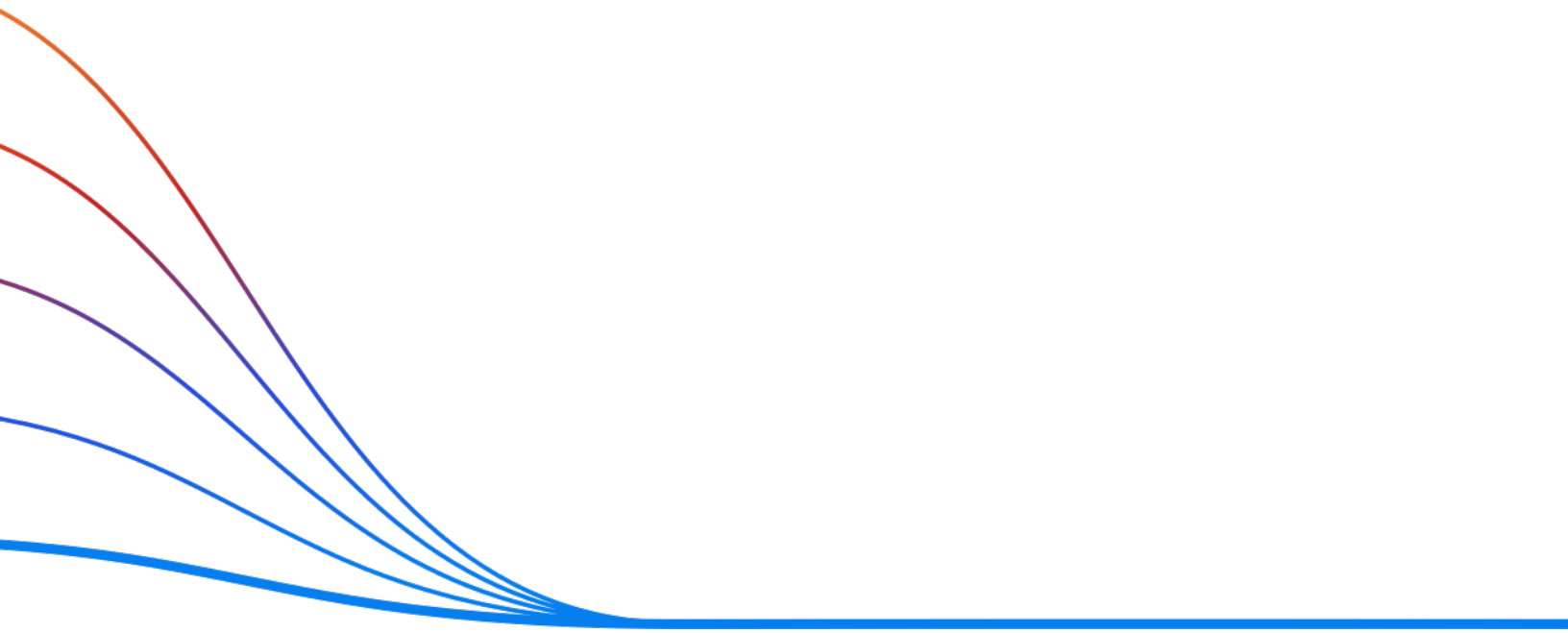
Figure 9: Residuals in a difference of differences calculation. Top - sample of residential meters with low evening ramp vs. a random sample of residential customers. Bottom - sample of residential meters with low evening ramp vs. a sample of residential customers that have similar load characteristics.

Importantly, the comparison group of randomly selected customers leads not only to higher variability in the COVID-period residuals, but these residuals exhibit a regular pattern of peaks every 24 hours that

would introduce upwards of 20% error in a load shape measurement *relative to total usage*. As a result of these and similar findings across several other trials detailed in Chapter 5 and Appendix B, random sampling should not be considered sufficient for the measurement of hourly load impacts. We revisit this topic in Chapter 3.

RECURVE
SHAPE THE FUTURE OF ENERGY

Chapter 3: Comparison Group Sampling Methods



I. Introduction

In the evaluation of demand-side energy programs, many comparison group sampling strategies have been developed and deployed. A full review of each approach is beyond this work, but common categories include random sampling, stratified sampling, future participants, and site-based matching. In developing recommendations for standardized methods to enable meter-based pay-for-performance programs, the following requirements are critical:

1. Methods need to enable non-participant sampling based on the population of treated customers.
2. Methods and required data must allow for tracking of a live program.
3. Methods must produce comparison groups amenable to statistical equivalence computations against the corresponding treatment groups.
4. Methods must result in as few subjective inputs and decisions as possible, even if that leads to greater complexity in the sampling execution and code.
5. Methods must not be prohibitively expensive to implement.

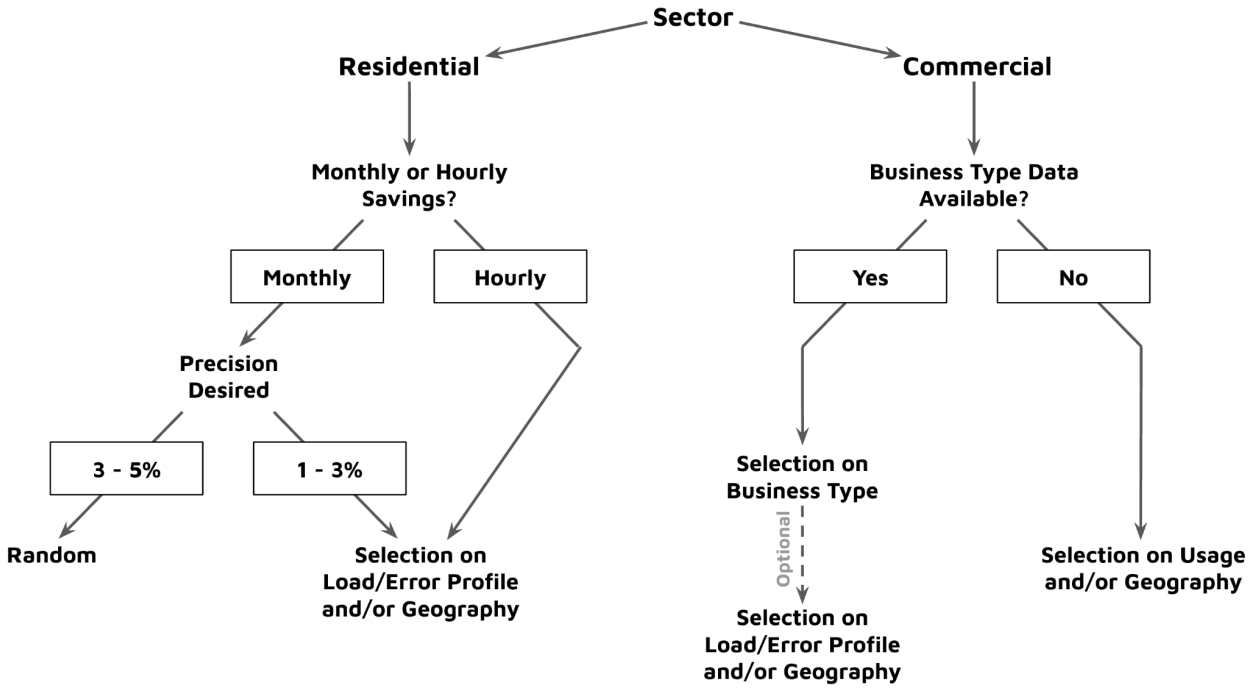
In addition to these conditions, the following recommendations are rooted in the testing of many different comparison group scenarios, which are detailed in Chapters 5 and 6 and complimentary appendices B and C.

While it can be desirable to have a single approach applied to all possible scenarios, our observations throughout the research and development phase of this effort have led us to an additional consideration:

6. Methods must be appropriate for the specific use case.

We have observed that while individual customers in the residential sector have exhibited a wide range of changes to energy consumption due to COVID, the *distributions* of meter-level COVID impacts have tended to be relatively consistent among different groups of customers. This contrasts with the commercial sector, where very clear and substantive differences are observed between different customer segments. As a result, when a measurement of total monthly or annual savings is the goal, random sampling may be a more appropriate and reliable approach for a residential program than for a program serving a specific commercial segment.

When hourly measurements or greater precision in total savings are needed, hourly measurements benefit from a comparison group selected based on additional criteria, including geographic location and usage characteristics (see Chapter 5/Appendix B). These results lead us to the following decision tree in the initial assessment of the type of sampling approach that should be employed for specific use cases:



II. Methodological Common Elements

In the next three sections we detail three comparison group sampling methods, Clustering, Individual Meter Matching, and Stratified Sampling that have been developed and tested across many cases. Stratified Sampling was the original GRIDmeter method, developed in 2020. In 2021 Recurve added Individual Meter Matching to this report and to the open source GRIDmeter repository. In this installment, Recurve has introduced Clustering. Each of these methods is appropriate for most measurement cases. However, Clustering is the most advanced method and is generally recommended.

Regardless of the exact sampling approach, there are many common elements to a comparison group-corrected savings calculation. Therefore, before detailing the specific sampling methods, in this section we detail the common elements that each method shares.

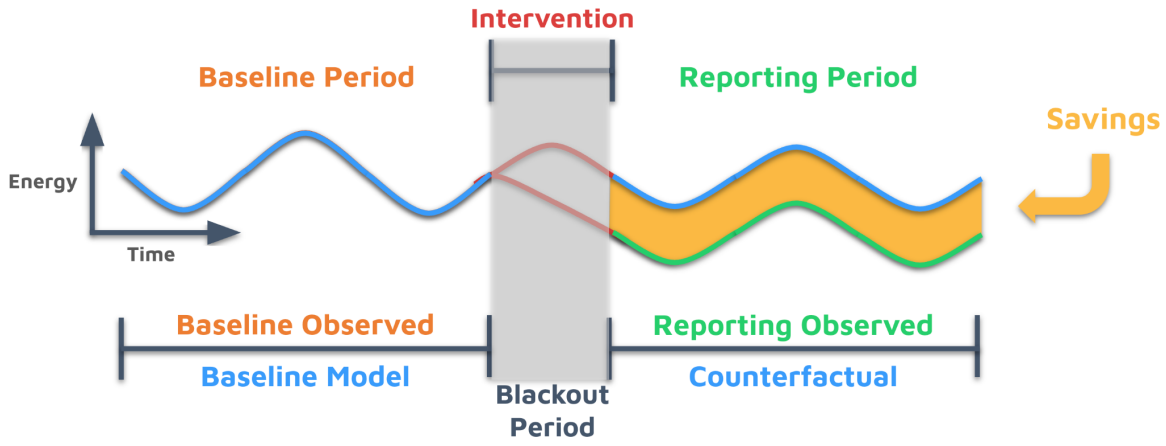
A. OpenEEmeter

All GRIDmeter comparison group sampling methods are based on open source OpenEEmeter baseline modeling as a first step. OpenEEmeter[5] methods were originally developed in partnership with the California Public Utilities Commission (CPUC), Pacific Gas and Electric Company (PG&E), and the California Energy Commission (CEC). OpenEEmeter specifies models for monthly, daily, or hourly consumption data and the specific model used depends on the temporal granularity of data provided. This report focuses on hourly electricity data and all baseline modeling was conducted with the OpenEEmeter 3.0 hourly model.

[5] The OpenEEmeter Python codebase, which runs the OpenEEmeter methods, is publicly available under the stewardship of the Linux Foundation Energy as an open-source project using a permissive Apache 2 license.

The OpenEEmeter 3.0 Hourly methods[6] use time-of-week and temperature variables. For each hour, usage is assigned across temperature bins (up to seven) and building occupancy state, which is estimated from a preliminary regression model. The temperature dependence is modeled through a piecewise linear function across the temperature bins.

The concepts of "baseline" and "reporting" periods are central to understanding the use of the OpenEEmeter model, comparison group selection, and load impact calculations. A schematic figure to illustrate these periods, along with a "blackout" period, is shown in Figure 1.



A schematic and timeline for savings calculations. The nomenclature defined here is used throughout the document.

The baseline period occurs before program intervention and a baseline model is built on weather and consumption data during that time. The reporting period occurs after the program implementation and during this period the baseline model is projected forward, using corresponding temperature data, as a counterfactual. Load impacts are calculated as the difference between the counterfactual and the actual consumption data during the reporting period. The timeframe when a change or intervention is in the process of being implemented within a building is known as the blackout period. This transient period does not represent energy usage patterns of either the baseline or reporting period and is therefore excluded from both fitting the model and when calculating savings.

For energy efficiency measurements, OpenEEmeter 3.0 Hourly requires that the baseline period must have at least 90% of days (or months) with non-zero, non-null values over one year. OpenEEmeter hourly methods have an additional requirement that there must be 90% coverage of each month. Event-based programs, such as demand response, can have much shorter baseline timeframes[7] and we discuss this further below.

[6] All methodological details can be found at https://github.com/openeemeter/eemeter.
[7] *Demand Response Advanced Measurement Methodology: Analysis of Open-Source Baseline and Comparison Group Methods to Enable CAISO Demand Response Resource Performance Evaluation*

B. Preliminary Segmentation

Before selecting a comparison group, a critical first step is the segmentation of both the treatment group and comparison pool into subgroups of meters that would be expected to experience similar trends on account of external factors. This step is done based on customer-level categorical information. Residential customers are initially segmented by at least climate region and solar panel status (true/false). In some cases, additional segmentation may be warranted, for example, by electric rate code if measuring an EV charge shifting program. For this reason, GRIDmeter methods are not explicitly prescriptive on categorical binning schemes beyond climate region and solar PV status in the residential sector.

However, research detailed in Chapter 6 and Appendix C shows that different business types exhibit vastly different changes in usage patterns due to the pandemic. Therefore, for long-term baseline scenarios (discussed in Section 1.3), it is best to segment commercial customers based on business type, in addition to climate zone and solar status.

Consider a commercial energy efficiency program serving hotels and grocery stores across two distinct climate regions. Every combination of solar status, business type, and climate region should be binned as an independent treatment group with comparison group selection occurring only from the comparison pool meters that share all of these attributes. In this case, there would be eight such subgroups. With the initial segmentation complete, comparison group selection based on features from time series consumption data can proceed as described in the sections below.

C. Baseline Period and Basis for Comparison Group Matching

Demand-side programs encompass measures expected to have load impacts over many years as well as interventions that influence usage for only a few hours, such as a demand response event. In the former, including energy efficiency projects, consumption patterns can be expected to change permanently, and it is important to measure savings over at least a full year. In the latter, consumption patterns are expected to return to normal shortly after the intervention period, and it may only be necessary to measure load impacts over a few weeks or even a single day. These distinct types of interventions can be cataloged into two general measurement scenarios, which we will refer to as long-term and short-term.

In the short-term scenarios, discussed in detail in Glass et al.,[7] interventions often occur on hot days, with the goal of reducing peak demand, and it is most important that the counterfactual is informed by a baseline model built from recent days with similar weather patterns. Comparison group selection therefore need not focus on distant months and should instead be focused on a relatively short window of time immediately preceding and following an intervention. In the long-term scenario, it is important that a measurement capture impacts during all seasons of the year and comparison group selection should be conducted accordingly. In both short-term and long-term scenarios, it is also important that comparison group selection and resulting counterfactuals are “aware” of differences between weekday and weekend patterns.

With these considerations in mind, when hourly data are available, a 168-point (7 days x 24 hours) average weekly load profile of a 45-day pre-intervention baseline period to perform comparison group selection for is used short-term cases. In long-term cases a 504-point seasonal hour-of-week profile, which consists of three concatenated 168-point average weekly load profiles for summer (June – September), shoulder (March – May and October), and winter (January, February, November, and December) timeframes is used. (The specified months are for California and can shift for other geographical regions.) When only daily usage data are available, these aggregations are not performed, and the daily data are used in comparison group selection.

D. Vintaging

In demand-side programs, customers are enrolled and interventions are completed over time. Recurve has measured programs that have enrolled customers over the course of five years or more. Therefore, comparison group selection and baseline/counterfactual generation should not take place over the same period for all savings calculations. Instead, GRIDmeter requires monthly vintaging in which every month's new program enrollees are matched to a distinct comparison group. Then the absolute %-difference-of-differences calculations, detailed in Chapter 4, are performed on OpenEEmeter models (baseline and reporting periods) that align with the enrollment month. In this way, if a program enrolls customers for a year, there will be 12 vintages for comparison-corrected savings calculations. Vintaging to achieve alignment of relevant baseline and reporting periods between treatment and comparison groups is a critical aspect of conducting accurate measurements in real-world applications.

E. Granularity and Aggregation

Both the treatment group and comparison group OpenEEmeter calculations are conducted at an individual meter level and produce outputs at the granularity of the raw meter data (down to hourly). Independent comparison group corrections are calculated for each unit of time (hour, day, etc.) at the comparison group level. Ultimately, comparison group corrected load impacts are calculated for each treatment meter (in the case of hourly consumption data). However, the comparison group correction itself should be thought of as a population trend correction as opposed to a correction specifically designed and computed to represent the behavior and circumstances of an individual participant. In other words, the goal of the comparison group is to accurately correct the portfolio-level savings; individual meter results should be understood with the caveat that the comparison group correction cannot resolve all specific non-programmatic effects that are encountered at an individual-building level.

Despite this internal incongruity, in the majority of use cases the most important result of load impact measurements is at the portfolio level. Fortunately, aggregation of savings results is straightforward via summation of corrected savings over the timeframe of interest across all participating meters.

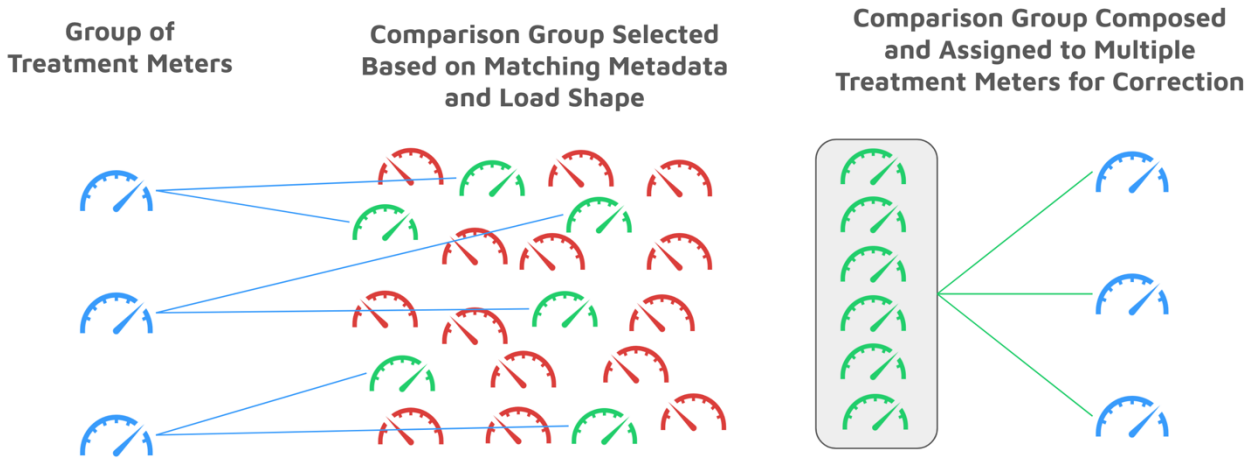
III. Individual Meter Matching

In the Individual Meter Matching sampling method, each treatment meter is matched with meters from the comparison pool that have the most similar aggregated load shapes. Individual Meter Matching

usually requires a comparison pool to be both large (thousands of meters if possible) and several factors larger than the treatment group (an order of magnitude being preferable) to operate optimally.

In Individual Meter Matching, a computation is performed on every combination of treatment and comparison pool meters to determine the similarity between each pair.[8] For an individual treatment/comparison meter pair this is done by calculating the Euclidian distance between a treatment and comparison pool meter aggregated load shapes.[9] This process is repeated without replacement until a comparison group of the desired size is constructed. The method is designed to be customizable, with the user selecting how many individual matches are desired for each meter. The comparison group is then formulated as all meters selected from the comparison pool and is used to correct the load impact calculations of all treatment meters at a population level. This process is shown in the schematic below.

Individual Meter Matching Conceptual Overview



The number of comparison meters matched per treatment meter should depend on the total number of treatment meters. As shown in Chapter 2, the residual error on account of the comparison group depends on the size of the sample.

At the end of the Clustering section, we show how Individual Meter Matching and Clustering compare across many samples.

[8] Or site aggregation

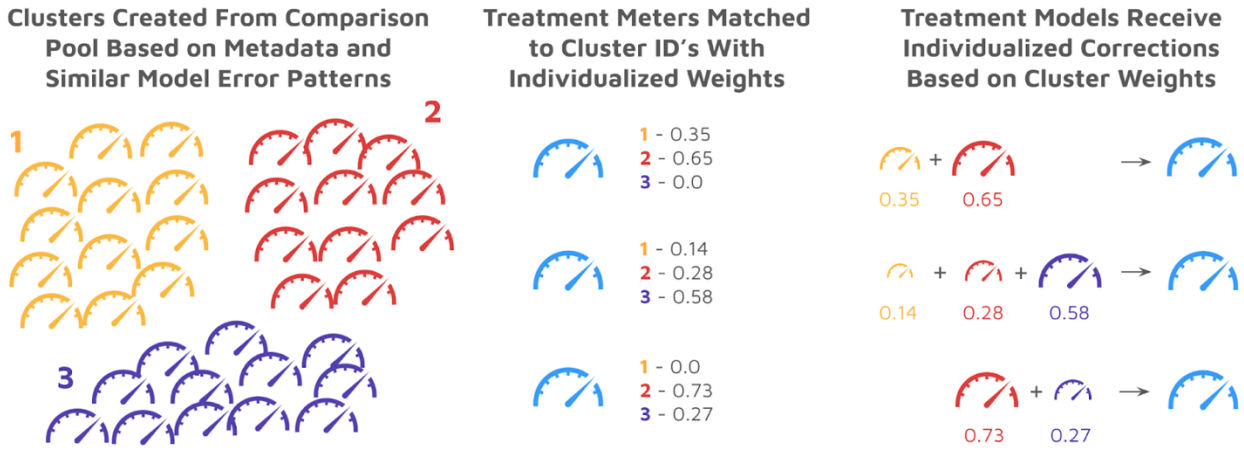
[9] In other words taking the square of the difference between treatment and comparison consumption for each month, summing these values and taking the square root of the total.

IV. Clustering

Clustering has been introduced as part of GRIDmeter 2.0. Through extensive testing Recurve has observed that it outperforms Individual Meter Matching (IMM) and we generally recommend its use where possible. Results of the full development and testing process are the subject of an academic paper that is under peer review as of the time of this writing. In this section we focus on defining the optimal method and sharing key results of testing done on the samples of residential and commercial customers described above.

The Clustering method consists of data transformation and machine learning steps to create groups or "clusters" of non-participant meters with similar model error profiles. Each treatment meter is assigned an individualized correction based on a weighted assignment of clusters. The concept of clustering is illustrated in the schematic below.

Clustering Conceptual Overview



The Clustering approach has several key advantages:

1. **Clustering harnesses more of the statistical power of the full non-participant pool**. With all suitable meters selected into a cluster, the comparison group takes advantage of the maximum statistical power available in the comparison pool. This trait of Clustering also mitigates the need for an evaluator to balance the size of the comparison group against the need for accurate load-shape matching, a tricky step that often requires judgment calls and is limited by the number of non-participants available.
2. **Clusters provide more stable and individualized comparison group corrections.** In Clustering, every participant is given an individualized correction that depends on the trends observed in the unique weighted combination of clusters assigned to it. The individualized correction is preferable for additional analyses that utilize meter-level results. In other comparison group methods, changing the composition of the participant sample will change the composition of a comparison group. In Individual Meter Matching, for example, adding a participant will add non-participant meters to the comparison group, thus changing the project results.

3. **Clusters are formed based on model error profiles, which are directly aligned with the goal of a comparison group.** Comparison groups are intended to isolate and remove model errors. By formulating clusters based on baseline period model error profiles, matching can then be done on the metric that best reflects the ability of a comparison group to accomplish its objective. For instance, a participant who experienced a 25% reduction in consumption during the COVID period will exhibit a model error profile that reflects this change. That participant will be matched to clusters that best emulate that pattern. Many other selection methods focus on consumption or load shape matching, which are often correlated but ancillary to the main objective of the comparison group.

Through extensive testing, Recurve has optimized the Clustering comparison group method. Below we describe the details of the approach and share key results from testing on residential samples.

The general procedure for creating comparison groups using clustering is:

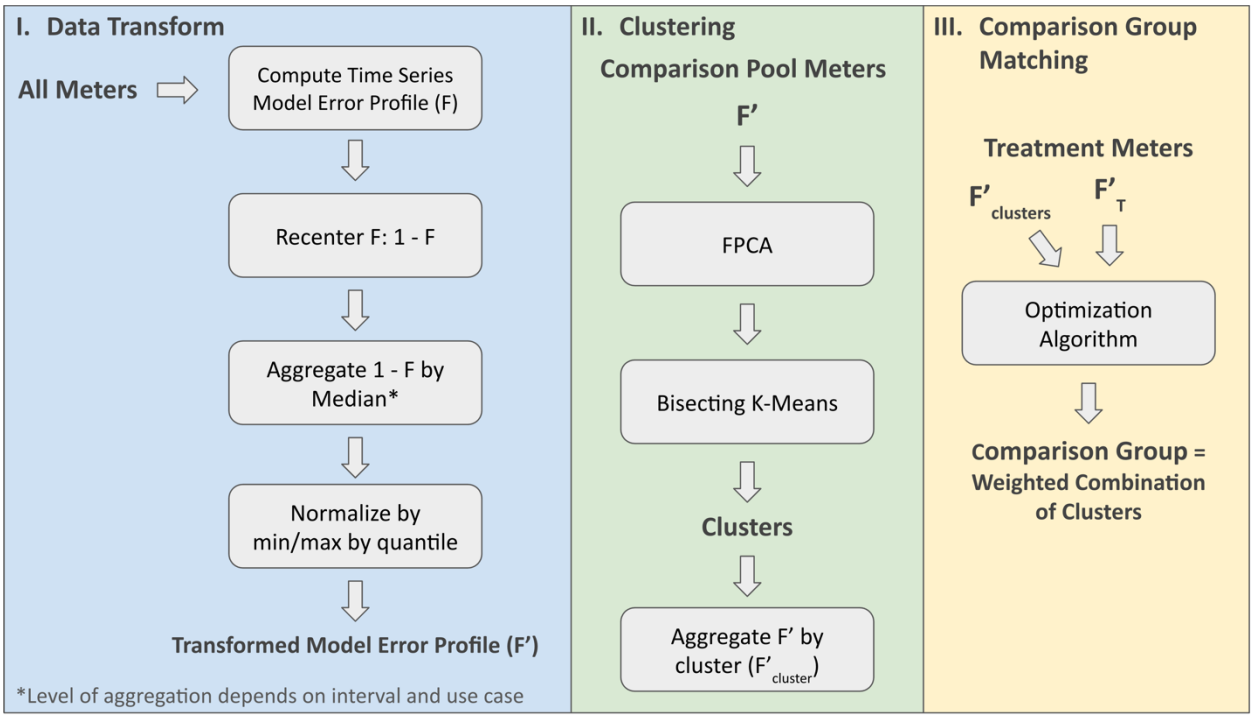
1. Create meter-level aggregated baseline model error profiles
2. Normalize model error profiles
3. Perform Functional Principal Component Analysis (FPCA) on model error profiles
4. Cluster the comparison pool FPCA outputs using the Bisecting K-Means Algorithm
5. Create comparison groups by assignment of each treatment meter to the optimal weighted combination of clusters

Important for the discussion that follows is Eq. 1, in which F is a dimensionless metric that gauges the degree to which the model is over or underpredicting consumption.

$$F = \frac{observed}{model} \tag{1}$$

The time series trace of this metric is the raw "model error profile" and is a key concept for clustering comparison group selection that is described below.

The following schematic outlines the full comparison group creation process that is described in detail in the following sections. This graphic can be a helpful reference to place any step in the overall method.



A. Match Basis

We begin the discussion of the clustering approach by revisiting the basis on which the similarity between treatment and comparison meters is gauged. Using consumption-based load shapes to assess similarity, as in Individual Meter Matching, is an intuitive method to formulate a comparison group. However, load shape is ancillary to the purpose of a comparison group, which is ultimately to remove model error. In Clustering we match on model error profiles directly. In this way, comparison group selection is conducted on the baseline period information that best gauges the expected ability of the comparison group to remove treatment group model error in the reporting period.

With baseline model error profiles calculated for every comparison pool meter, we can establish groups (or "clusters") of meters that all exhibit similar error patterns. By then matching each treatment meter to the weighted combinations of comparison group clusters that best recreates the treatment meter's model error profile, we reduce systematic errors by effectively dividing them out. The same concept applies in Individual Meter Matching but supposes a strong correlation between consumption load shape and model error profile. While there are often good reasons to expect this, it is not necessarily a given and the degree can certainly vary.

A good example of a model error pattern emerges when considering seasonal bias, a common phenomenon in which a model under or overpredicts consumption depending on the time of year. By matching on model error profile, when a treatment meter exhibits seasonal bias, those same patterns present in comparison pool meters will lead to selection into the comparison group. The assumption, aided by the categorical binning described above in section II.B, is that the model error patterns will continue to match between treatment and comparison group during the reporting period, at which

point the treatment counterfactual is "corrected" by removing the model error patterns observed in the comparison group per the difference of differences methodology described in Chapter 4.

While it would be possible to generate clusters directly from raw model error profiles, additional data transformations are very helpful to promote computational stability, limit computational cost, and to optimize cluster generation and treatment-meter matching. Much of the discussion that follows details the data transformation steps undertaken to maximize performance of the clustering methods.

B. Transformations of Model Error Profiles

i. Aggregation and Recentering

Hourly energy time series data can be highly variable and using raw traces for complex calculations can be very expensive. Therefore, as described above, it is beneficial to generate aggregated traces that still enable the capturing of important time series characteristics. With hourly data, the operation to aggregate to 504-point seasonal hour-of-week traces for the long-term baseline or 168-point hour-of-week traces for the short-term baseline is generally performed by combining all hours in the same period together. For example, all summer Mondays at 10 AM are combined into a single point on the seasonal hour-of-week curve. After testing multiple aggregation functions, optimal performance is observed utilizing median values. Therefore, for the clustering method, we aggregate the model error profiles, $1 - F$, utilizing the median value. Using the summer Monday, 10 AM example, 17 such hours will contribute to the 504-point aggregation and the median value of these 17 is taken.

When daily or monthly data are available, the recentered 365-point or 12-point baseline model error profiles can be used without further aggregation.

ii. Normalization

In making comparison group corrections, normalization of the model error profiles can help to clearly isolate trends and features, which will serve as the basis for differentiation when forming clusters of similar meters. After researching many possible normalization algorithms, we have identified the minimum/maximum by quantile (Eq. 2) as the best performing option.

$$y_i = \frac{x_i - Q_{0.1}(x)}{Q_{0.9}(x) - Q_{0.1}(x)} \quad (2)$$

In Eq 2, x_i is a specific point in the aggregated model error profile (median value of summer, Monday, 10 AM for example), y_i is the corresponding normalized point, and $Q_{0.1}$ and $Q_{0.9}$ are the 10th and 90th percentile values of the full aggregated model error profile (x). It is important to reemphasize that these transforms are not being applied to reporting data, but are only utilized for the purpose of creating comparison groups.

Figure 10 shows the components of the data transformations for a single meter (this will be taken as a "treatment" meter to illustrate further steps described below). The top panel shows the median

baseline period observed (solid blue) and model (dashed orange) values aggregated to the 504-point seasonal hour-of-week basis. The solid red trace of the bottom panel results from applying $1 - F$, to each point. Finally, the bottom panel shows the normalized error (dashed green), which results from the subsequent application of Eq. 2. This "transformed error" trace will be the subject of further analysis described below.

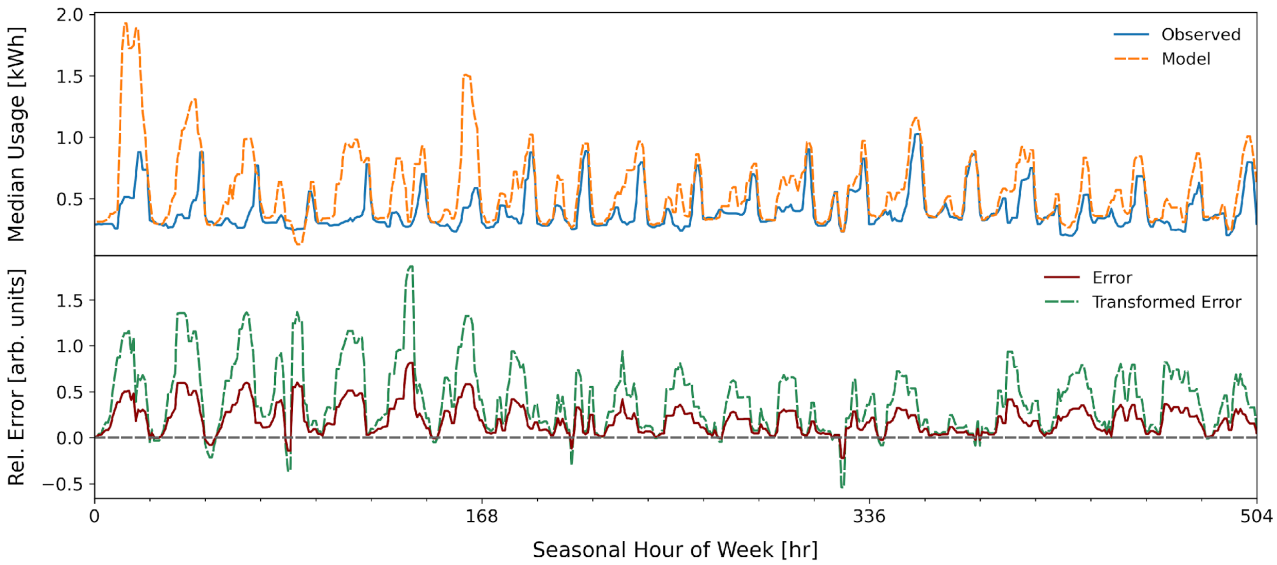


Figure 10. Seasonal hour-of-week profiles for: top – median observed load shape and model; bottom – the model error profile resulting from application of 1 – F and the transformed error profile that subsequently results from normalization via Eq. 2.

iii. Functional Principal Component Analysis (FPCA)

While the aggregation step of the previous section greatly reduces hourly traces, it is still not advisable to operate clustering algorithms on vectors with hundreds of points. Therefore, a functional principal component analysis (FPCA) is used to further reduce the model error profiles to features.[10] In an FPCA, a linear combination of component (basis) functions is used to emulate a trace. All meters use the same set of basis functions with the coefficients themselves encoding the information needed to characterize the model error profile of a given meter and subsequently form clusters.

[10] An FPCA bypasses the common concern that a PCA applied to time-series data ignores the temporal dependence each data point has on those surrounding it. There are three common methods of clustering time-series data: raw-data-based, feature-based, and model-based. The model-based approach tends to be computationally intensive and hence untenable given the large datasets routinely analyzed in demand side programs. In the raw-data-based approach the transformed model error profiles are used directly in the clustering algorithm. However, clustering tends to perform better with fewer variables. Therefore, we adopt the feature-based approach to formulate clusters. See: T.W. Liao, Clustering of time series data—a survey, Pattern Recognition. 38 (2005) 1857–1874. https://doi.org/10.1016/j.patcog.2005.01.025.

As a trivial example, consider two basis functions, a cosine wave (C) and a square wave (Sq). One meter may be best represented by 0.1 x C + 0.9 x Sq while another by 0.8 x C + 0.2 x Sq. The vector of these coefficients for all meters in a sample can then be passed to a clustering algorithm.

Two variables must be defined for the FPCA implementation:[11] the functional basis for creating the basis expansion and the minimum explained variance ratio.[12] A Fourier basis is used since it performs well given the regular, repeating patterns often observed in these traces. The minimum explained variance ratio, optimized to 97% in this study, establishes the fidelity to which the component functions must recreate the variance of the transformed model error profiles.

Returning to our example that began with Fig. 10, Fig. 11 shows the FPCA process. The thin blue trace of the top panel of Fig. 11 replicates the transformed model error profile (the dashed green curve of Fig. 10). The thick black trace is the representation of this curve achieved through the combination of FPCA components. The bottom panel shows the first three Fourier basis components (out of the 64 needed to achieve the 97% minimum explained variance threshold).

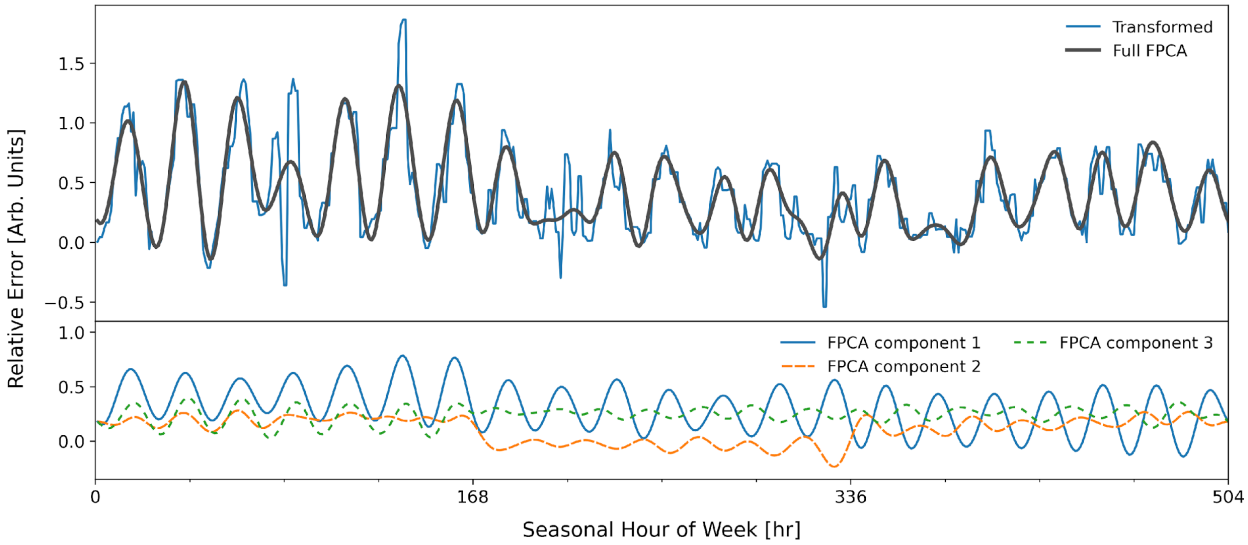


Figure 11. Top: The transformed model error profile (thin blue line) and results of the FPCA expansion (thick black line) for the example meter. Bottom: The most prominent 3 (out of 64 total) FPCA components that when summed yield the thick black line of the top panel.

[11] Implementation of FPCA is done through the Scikit-fda Python package.

[12] The minimum explained variance ratio is an optimization parameter with a minimum value of 50% and a maximum of 100% that defines how well the basis function expansion needs to replicate the trace. After optimizing the algorithm, we impart a 97% minimum explained variance. In other words, additional basis functions will be added to the FPCA linear combination until the resulting curve matches well enough to meet a threshold of 3% or less variance between the two curves.

C. Bisecting K-means Clustering

With the vectors of FPCA coefficients as inputs, sklearn's bisecting K-means clustering algorithm is used to generate clusters of comparison pool meters with similar model error patterns.[13] The bisecting K-means[14] algorithm was the best performing of several tested, including the BIRCH algorithm, K-nearest neighbors, and other variants of K-means.

i. Cluster Scoring

Evaluation of the clusters utilizes the Variance Ratio Criterion, which was selected from several scoring algorithms tested as part of this research as it produced the highest performing comparison groups.

ii. Cluster Sizing and Minimum/Maximum Number of Clusters

Bisecting K-means operates off a set number of clusters, a parameter that is input prior to running the algorithm. This is challenging because it is impossible to know the optimal number of clusters *a priori*. To address this, each number of clusters is tested up to a maximum and the score of the clusters is calculated for each number. The number with the best score is selected as optimal. The maximum number of clusters checked varies with the comparison pool size but will never exceed 1500 clusters.

The clusters are ultimately used to reduce and correct bias and uncertainty in treatment models. To best accomplish this, it is important to balance cluster size and the similarity of meters within a cluster. Very large clusters will tend to provide high measurement stability but will contain a wide variety of meters, which can sacrifice accuracy. To safeguard against clusters that are too small, a minimum meter threshold of 15 was determined to be optimal through testing. If the number of meters in a cluster falls below this threshold, then the cluster is rejected.

In some applications the number of comparison pool meters may be small, on the order of a few dozen or less, and in others the comparison pool can be very large. Given this, it is helpful to define minimum and maximum number of clusters that vary by the size of the comparison pool. The equations to define these minimum and maximum values are somewhat complex and we refer interested readers to the functions in the open-source codebase.

D. Comparison Group Assignment

Once the final clusters have been defined, the last step in formulating comparison groups is the assignment of each treatment meter to clusters. Every treatment meter is assigned to a *mixture* of

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research. 12 (2011) 2825–2830.

[14] In bisecting K-means the largest cluster is bisected into 2 clusters using K-means. In this context largest can refer to either the cluster with the biggest SSE or with the largest number of data points. This breaks the problem down into more manageable chunks because each successive split is done on a smaller and smaller dataset, requiring less computational expense.

clusters because a treatment meter's transformed model error profile can often exhibit distinct features that may be more or less prominent in a given cluster.[15] Matches to clusters are made based on the Euclidean distance between the treatment meter and the clusters' transformed model error profiles. Based on this metric, the most similar cluster will be given the greatest weight and the least similar cluster the least weight (or no weight).

A given cluster will be composed of some number of meters, each with its own transformed model error profile. This raises the question of what transformed model error profile should be taken to represent the cluster. Here again we rely on a transformed median approach. Within a cluster we compute the median transformed model error profile and pass that through the min/max by quantile normalization of Eq. 2. The resulting trace is then taken as the cluster's transformed model error profile for the purpose of assigning weights during the matching step.

To speed the optimization process and bypass the need for a global optimization scheme, the initial guess for the cluster weights is calculated using the distances of the treatment meter from the clusters as described in Eqs. 3 – 5.

$$d_i = distance(F'_T, F'_{cluster}) \tag{3}$$

$$w_{0,i} = \left(\frac{(d)}{d_i}\right)^{20} \tag{4}$$

$$w = \frac{w_0}{sum(w_0)} \tag{5}$$

In Eqs. 3, d_i is the Euclidian distance between the transformed treatment meter and a cluster i's model error profiles (F'_T and $F'_{cluster}$ where the $'$ indicates the model error profile has been aggregated, recentered, and normalized as described above). In Eq. 4, non-normalized weights, $w_{0,i}$, are computed by taking the 20th power of the minimum of all d_i ($min\ (d)$) divided by d_i (this quantity equals 1 for the value of d_i corresponding to the minimum). The set, w, results from normalizing the set of w_0 to 1 and provides the initial guesses for the weights being optimized. The power of 20 in Eq. 4 was determined based on preliminary global optimization tests.[16]

[15] Examples of error patterns include seasonal and weekday/weekend biases. For example, a meter can show large positive error during summer and negative error during the weekends but only a combination of clusters can represent this.

[16] The optimization methodology described works adequately, but there is room for refinement in future work. Since the optimization does not include a global element, it is possible that it could get stuck in local minima and not return the optimal cluster weights. Because the initial weights are determined by distances, it is possible that the best set of weights could include a small percentage of a cluster that does not fit well in order to capture some behavior from it. Despite this concern, through testing it was found that the current methodology produces well-fitting comparison groups.

For each treatment meter, the final weights assigned to each cluster are determined using sequential least squares programming in conjunction with Barron's adaptive loss function and suggestions from Chebrolu et al.[17,18] Two constraints are imposed on the weights: they must be positive, so clusters cannot be subtracted, and they must sum to 1.

Continuing with our example, we return to the transformed model error profile of Fig. 10, which is again replicated as the thin blue line in the bottom panel of Fig. 12. The transformed model error profiles of the four primary matched clusters are shown in the top four panels with the percentages in the top right corners corresponding to their weightings. The weighted sum of the cluster model error profiles yields the thick black trace in the bottom panel. The weighted sum of clusters clearly reproduces most of the core patterns of the treatment meter. The "daylight" between curves is expected. Any individual treatment meter will exhibit noise and fluctuations. When aggregated across an entire portfolio of treatment meters, the match between treatment and comparison groups will be much stronger.

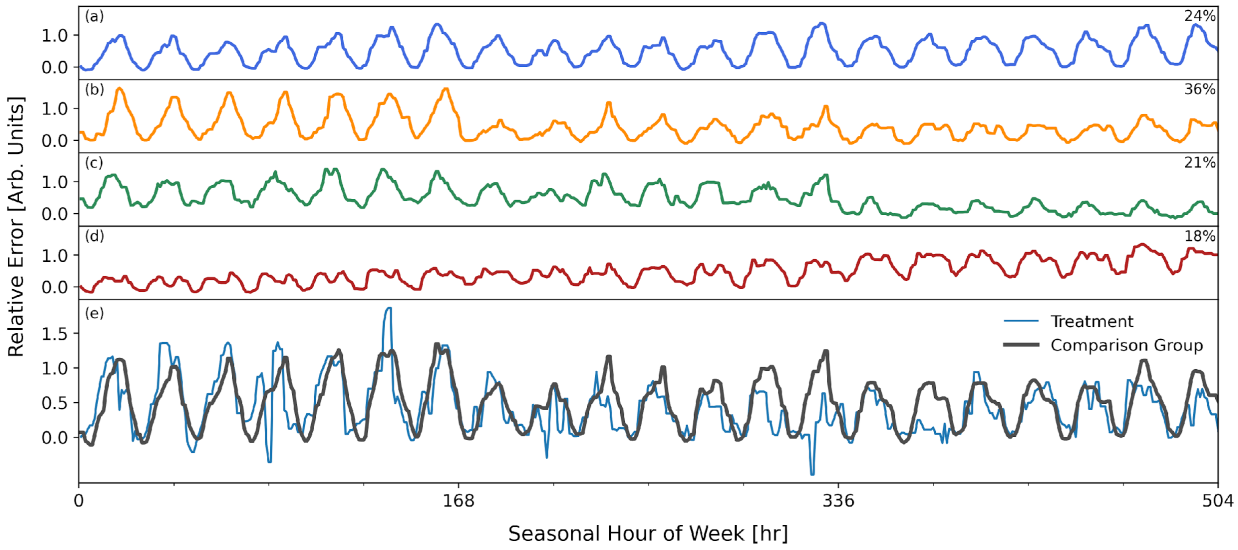


Figure 12. The transformed model error profiles for the four clusters that the example treatment meter maps to (a – d) and the composite comparison group compared to the treatment meter (e). Combining the 4 clusters using the percentage in the top right corner of each panel accounts for 99% of the treatment meter's comparison group in (e).

Figure 12 is a good example of how different features from clusters can be linearly combined to create a comparison group that best matches the transformed model error profile of the treatment meter. Many features in Figure 12b, such as the regular triangular patterns in the summer (hours 0 – 168), are prominent in the comparison group transformed model error profile, Figure 12e, but other clusters are needed to represent the shoulder months (hours 168 – 336).

[17] J.T. Barron, A general and adaptive robust loss function, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: pp. 4331–4339.

[18] N. Chebrolu, T. Läbe, O. Vysotska, J. Behley, C. Stachniss, Adaptive Robust Kernels for Non-Linear Least Squares Problems, IEEE Robotics and Automation Letters. 6 (2021) 2240–2247. https://doi.org/10.1109/LRA.2021.3061331.

E. Summary of Parameters

Table 2 gives the optimized values or approaches for the key parameters/strategies tested as part of the clustering work.

Table 2. A complete list of the Clustering method variables and their optimized parameters.

Variable	Optimized Parameters
Clustering Algorithm	Bisecting K-means
Error Profile Aggregation	Median
Error Profile Normalization	Min/max by Quantile
FPCA minimum explained variance	0.97
Scoring algorithm	Variance Ratio
Minimum cluster size	15
Distance metric	Euclidean

The next two sections define uncertainty and the objective function against which the parameters of Table 2 were optimized.

F. Savings Uncertainty

While comparison groups decrease the total error by accounting for factors outside of the model inputs, they are not without cost. Comparison groups have their own uncertainty contribution to savings. Eq. 7, shows how to calculate the uncertainty for F_{CG}, and Eq. 8 shows how the comparison group correction factor uncertainty propagates into the savings uncertainty on a per-meter, per-hour basis. Full derivations of Eqs. 7 and 8 are provided in the Methods Appendix.

$$\epsilon_{F_{CG}} = \sqrt{\sum \left(w_i \left(\epsilon_{F_{CG_i}}^2 + \left(F_{CG_i} - F_{CG} \right)^2 \right) \right)} \frac{t_{\alpha/2,N-1}}{\sqrt{N}} \tag{7}$$

$$\epsilon_S = \sqrt{\left(\epsilon_{F_{CG}} m_T \right)^2 + \epsilon_{m_{T,c}}^2 + 2\, cov\left(F_{CG}, m_{T,c} \right) m_T} \tag{8}$$

where ϵ is uncertainty, F_{cg} is the "comparison group correction factor" ($\frac{o_{CG}}{m_{CG}}$), w is the weight, i indicates that the quantity is per meter, t is the t-value, N is the number of samples being averaged, m_{CG} is the comparison group model, m_T is the uncorrected treatment model, $m_{T,c}$ is the corrected treatment model, S is the savings, and cov is covariance between F_{CG} and $m_{T,c}$. Eq. 7, uses a prediction interval to estimate the uncertainty of the F_{CG}. When utilizing the new comparison group methods, the weights in Eq. 7 can be neglected but when calculating uncertainty for Individual Meter Matching those weights are the percent of total consumption for each meter within the comparison group for a given hour. A prediction interval is more appropriate than a confidence interval given the goal is to predict the range a treatment meter's F_{CG} might lie within in the absence of a program intervention.

The savings uncertainty, Eq. 8, is a combination of uncertainties from F_{CG} and m_T, in other words the final savings uncertainty accounts for the uncertainty of the treatment model and the uncertainty introduced by the comparison group correction. The uncertainty due to the covariance of these two variables is also included. Since these equations are calculated on a per-meter, per-hour basis, and savings are usually reported for aggregations along one or both dimensions, the proper way to aggregate these uncertainties is to add them in quadrature for the hours/meters being summed.

Taking a step back, the overall uncertainty might appear to be increasing with the use of a comparison group because the comparison group itself is an uncertain quantity. However, we are in fact trading unknown, unquantified uncertainty for quantified uncertainty. For example, in the instance of COVID-19, a building might show relatively small uncertainty if no comparison group is used, but in fact, the entire result is biased and unreliable unless the impacts of COVID-19 are somehow considered in the model.

G. Objective Function

There are multiple ways to gauge the success of a comparison group (precision, accuracy, variance etc.) and we must determine whether to use one or some combination of metrics. We choose an objective function that minimizes a combination of bias error, which the comparison groups might introduce, and savings uncertainty. The objective function is explicitly provided as Eq. 9.

$$objective = 10\ \epsilon_{S_{bias}} + \sqrt{\sum \epsilon_{S_i}^2} \tag{9}$$

where $\epsilon_{S_{bias}}$ is bias error and ϵ_{S_i} is the savings uncertainty of an individual treatment meter. The uncertainties are aggregated across the reporting period and normalized by the total consumption of the meters within the sample to enable comparison of uncertainties between samples.

This objective function is used to optimize the hyperparameters and methodological choices within the data transformation and clustering steps described above and collected in Table 2. This objective function can also be applied to Individual Meter Matching for the purpose of comparison.

With the assumption that that aggregate savings is zero given the non-participant sample, bias error can be determined by computing the savings confidence interval and checking if that quantity overlaps with zero. If it does, then no bias error is assigned. Otherwise, the bias error is taken as the distance of the confidence interval from zero, normalized by the total savings, and scaled by a factor of 10. The factor of 10 heavily penalizes comparison groups that introduce systematic bias.

H. Samples Tested

Table 3 details the samples tested. In some cases, including those from the commercial sector, treatment and comparison pool samples are drawn randomly at a 1:10 ratio. In the bulk of residential cases, the treatment samples are purposefully selected to have different characteristics than the comparison pool. This is done to emulate cases in which a program serves a population of customers who are distinct from a random selection. For instance, a program that replaces air conditioners may

recruit customers with very high baseline period air conditioning loads and having an air conditioner may be a basic eligibility criterion for the program. Such cases are clearly indicated. For example, in the sample “Top 25% annual consumption,” the 100 treatment meters are pulled exclusively from residential customers in the top quartile of annual consumption among all residential customers. The 1000 associated comparison pool customers, however, remain a random sample, thus emulating a real-world comparison pool, which often will not be stratified to match the treatment sample.

Where at least 1100 meters were available for a given sample, we pulled 100 into a treatment group and 1000 into the comparison pool. Where fewer than 1100 meters were present in a given commercial subsector, we selected 10% into a treatment sample and 90% into a comparison pool. Overall, the samples were chosen to represent a wide variety of different usage patterns, variability within samples, and to represent many different potential use cases across different segments of the population and sectors of the economy.

Table 3: Descriptions of all 62 samples along with bias, fractional uncertainty, and objective function results for both Individual Meter Matching and Clustering methods.

			IMM			Bisecting K-Means		
Treatment	Comparison Pool	Sample	Bias	Unc.	Obj.	Bias	Unc.	Obj.
100	1000	Random	-	0.148	0.148	-	0.118	0.118
100	1000		-	0.155	0.155	-	0.123	0.123
100	1000		-	0.170	0.170	-	0.143	0.143
100	1000		-	0.170	0.170	-	0.140	0.140
100	1000		-	0.322	0.322	-	0.282	0.282
100	1000	Top 25% annual consumption	-	0.157	0.157	-	0.125	0.125
100	1000	Bottom 25% annual consumption	-	0.181	0.181	-	0.129	0.129
100	1000	Top 10% evening ramp (1)	-	0.178	0.178	-	0.138	0.138
100	1000	Bottom 10% evening ramp (1)	-	0.163	0.163	-	0.133	0.133
100	1000	Top 20% summer peak load (2)	-	0.157	0.157	-	0.118	0.118
100	1000	Bottom 20% summer peak load (2)	-	0.159	0.159	-	0.117	0.117
100	1000	Top 20% winter load (3)	-	0.145	0.145	-	0.122	0.122
100	1000	Bottom 20% winter load (3)	-	0.207	0.207	-	0.159	0.159
100	1000	Top 25% annual cons. and Top 50% of discretionary (4)	-	0.154	0.154	-	0.126	0.126
100	1000		-	0.151	0.151	-	0.122	0.122
100	1000		-	0.153	0.153	-	0.130	0.130
100	1000	Top 30% evening ramp1 and Top 25% of summer peak load (2)	-	0.162	0.162	-	0.127	0.127
100	1000		-	0.158	0.158	-	0.126	0.126
100	1000		-	0.165	0.165	-	0.123	0.123
100	1000	Top 50% cooling load6 and Top 25% of baseload (5)	-	0.178	0.178	-	0.142	0.142
100	1000		-	0.154	0.154	-	0.126	0.126
100	1000		-	0.172	0.172	-	0.138	0.138
100	1000	Top 50% cooling load6 and Bottom 25% of baseload (5)	-	0.225	0.225	-	0.159	0.159
100	1000		-	0.218	0.218	-	0.159	0.159
100	1000		-	0.224	0.224	-	0.159	0.159
100	1000	Bottom 25% of cooling6 and Bottom 25% of heating (7)	-	0.159	0.159	-	0.131	0.131
100	1000		-	0.190	0.190	-	0.175	0.175
100	1000		-	0.150	0.150	-	0.131	0.131
100	1000	Top 25% summer peak load2 and Bottom 50% shoulder (8)	-	0.174	0.174	-	0.152	0.152
100	1000		-	0.236	0.236	-	0.158	0.158
100	1000	Unassigned/Unknown business type	-	0.800	0.800	-	0.451	0.451
100	1000		-	0.259	0.259	-	0.208	0.208
100	1000		-	0.435	0.435	-	0.353	0.353
100	1000		-	0.515	0.515	-	0.570	0.570
100	1000	Real Estate/Leasing	-	0.179	0.179	-	0.158	0.158
100	1000		-	0.227	0.227	-	0.180	0.180
100	1000		-	0.193	0.193	-	0.176	0.176
100	1000	Administrative/Civil	-	0.051	0.051	-	0.060	0.060
100	1000		-	0.052	0.052	-	0.060	0.060
100	1000		0.017	0.068	0.233	-	0.065	0.065
100	1000	Cellular/Cable/Wireless	-	0.011	0.011	-	0.008	0.008
100	1000		-	0.009	0.009	-	0.008	0.008
100	1000	Medical/Veterinary Offices	-	0.211	0.211	-	0.168	0.168
100	1000		-	0.660	0.660	-	0.542	0.542
100	1000	Retail	-	0.302	0.302	-	0.246	0.246
100	1000	Offices	-	0.229	0.229	-	0.188	0.188
89	890	Automotive	-	0.141	0.141	-	0.128	0.128
86	860	Beauty	-	0.401	0.401	-	0.299	0.299
85	850	Construction/Contractors	-	0.507	0.507	-	0.517	0.517
79	790	Restaurants/Bars	-	0.559	0.559	-	0.432	0.432
55	550	Sports/Entertainment	-	1.530	1.530	-	1.134	1.134
51	510	Wholesale Trade	-	0.110	0.110	-	0.114	0.114
51	510	Churches/Religious	-	0.289	0.289	-	0.280	0.280
44	440	Warehousing/Shipment/Transportation	-	0.080	0.080	-	0.091	0.091
43	430	Hospitals/Care	-	0.386	0.386	-	0.393	0.393
41	410	Utilities	-	0.050	0.050	-	0.075	0.075
31	310	Agriculture/Farming	-	0.119	0.119	-	0.129	0.129
24	240	Hotels/Lodging	-	0.130	0.130	-	0.127	0.127
21	210	Schools	-	0.177	0.177	-	0.170	0.170
18	180	Grocery/Convenience	-	0.057	0.057	-	0.055	0.055
17	170	Laundry/Dry Cleaning	-	0.198	0.198	-	0.154	0.154
17	170	Banks	-	0.044	0.044	-	0.046	0.046

[1] Evening ramp: the difference in a meter's average hourly usage from 6 – 7 pm compared to 2 – 3 pm
[2] Summer peak load: the total usage from 4 – 9 pm during June – September
[3] Winter load: the total usage during November – February
[4] Discretionary: an estimate of a meter's non-temperature dependent, non-baseload usage
[5] Baseload: a meter's average non-temperature dependent, minimum consumption
[6] Cooling load: the percentage of a meter's annual usage found to be temperature dependent in warm weather
[7] Heating load: the percentage of a meter's annual usage found to be temperature dependent in cold weather
[8] Shoulder load: the total usage during March – May and October

I. Results of Testing Individual Meter Matching and Clustering

For each sample, comparison groups were generated as described above, using the optimal parameters of Table 2 for the Bisecting K-means approach. We then calculated reporting period savings per the difference of differences method described in Chapter 4. Because all meters are program non-participants,[19] it can be assumed that zero savings is the correct result within any sample. With this assumption, bias (non-zero savings) and uncertainty were measured for each sample.

Table 3 gives results for all 62 samples studied. Figures 13 (residential) and 14 (commercial) give the objective function results as a distribution of relative improvement compared to Individual Meter Matching for three candidate clustering methods originally considered. While we leave discussion of KNN and BIRCH to the academic literature, it is clear that Bisecting K-means was the best performing of the three and produced results significantly better than Individual Meter Matching in most cases.

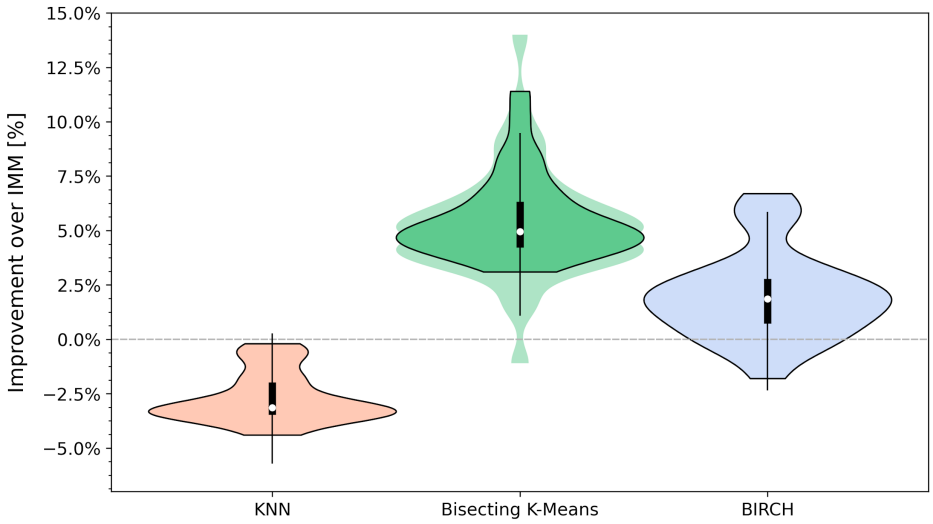


Figure 13. A violin plot of the percent improvement in the objective value of the three optimized algorithms compared to the Individual Meter Matching (IMM) method for residential customers. Positive values represent improvement over IMM. The thin black lines represent the 1.5 interquartile range, the solid black line is the interquartile range, and the white circle is the median of all 62 subsamples. The shape represents the distribution of values using kernel density estimation. The shading around the bisecting K-means algorithm represents the minimum and maximum results of 22 repeats. The median improvements are 3.2% (KNN), 5.0% (Bisecting K-Means), and 1.9% (BIRCH).

[19] More specifically, no program information is being used to filter between groups

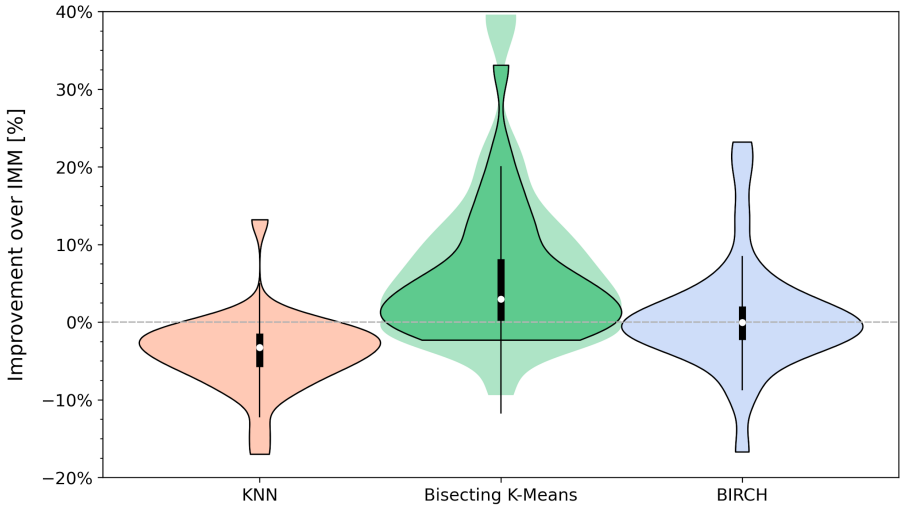


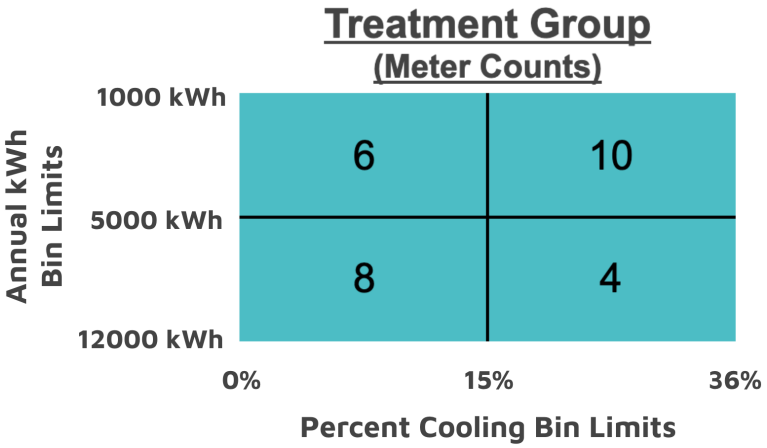
Figure 14. A violin plot of the percent improvement in the objective value of the three optimized algorithms compared to the prior Individual Meter Matching (IMM) method for commercial customers. Positive values represent improvement over IMM. The thin black lines represent the 1.5 interquartile range, the solid black line is the interquartile range, and the white circle is the median of all 62 subsamples. The shape represents the distribution of values using kernel density estimation. The shading around the bisecting K-means algorithm represents the minimum and maximum results of 22 repeats. The median improvements are -3.3% (KNN), 3.0% (Bisecting K-Means), and -0.1% (BIRCH).

V. Stratified Sampling

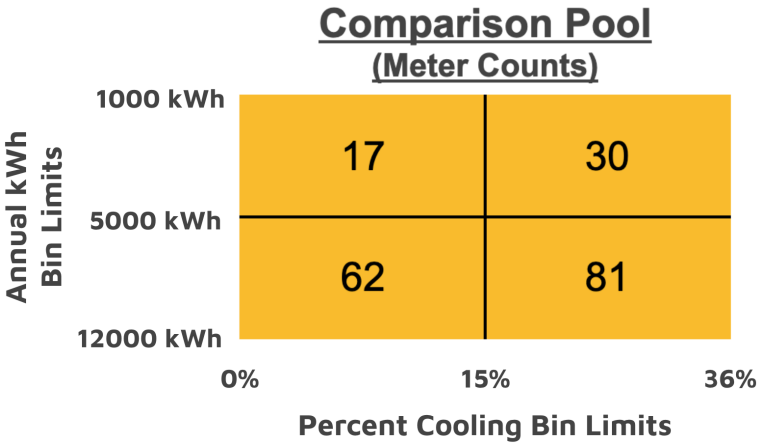
In stratified sampling, a comparison group is generated from a comparison pool by selectively eliminating members of the comparison pool until the remaining sample is representative of a treatment group. This elimination step is done based on one or more characteristics that can be known or measured for all members of both treatment and comparison samples. In practice, stratification is done by identifying specific, quantifiable parameters of interest and then forming discrete bins based on the values of those parameters observed in the treatment group. Each customer is assigned to a single bin. The resulting counts within each bin determines the proportionality that must then be matched by sampling from the comparison pool.

A. Illustrative Example: Traditional Stratified Sampling

To illustrate the concepts of multidimensional stratified sampling based on usage characteristics, we consider the following treatment group and comparison pool where stratification is to be done based on the parameters of total annual kWh usage (annual_kwh) and the percentage of a customer's usage from cooling (pct_cooling). To keep the example tractable, we split each parameter into just two bins, creating a total of four bins. The figure below shows how meters in the treatment group are assigned to bins given the indicated limits. A meter with 7,430 annual kWh and 11% cooling would be one of the 8 meters assigned to the lower-left bin.



With the same bin limits the comparison pool has the following meter counts:



Comparing the fraction of meters in each bin, we see that the comparison pool differs from the treatment group:

Treatment Group (Fraction of Meters)	
0.21	0.36
0.29	0.14

Comparison Pool (Fraction of Meters)	
0.09	0.16
0.33	0.43

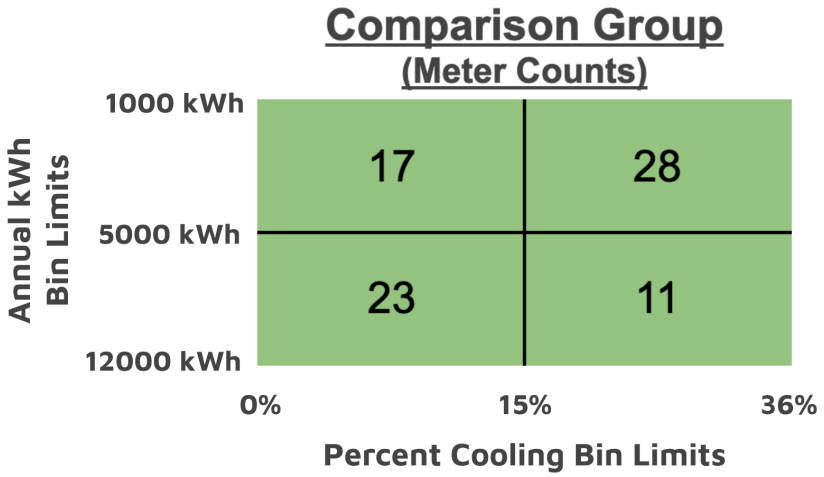
The job of stratified sampling is to create fractional parity between the treatment and comparison groups within each bin. Since the comparison pool is set, meters cannot be added to underrepresented bins. Instead, meters must be eliminated from bins that are oversampled compared to the treatment group. For any binning scheme, a "limiting bin" will emerge corresponding to the most undersampled bin in the comparison pool. The limiting bin is determined by taking the ratio of population in the comparison pool vs. treatment group for each bin and locating the minimum. In the current example, this step results in the following:

Comparison vs. Treatment (Population Ratio)	
0.42	0.44
1.14	2.98

With the upper left bin having the lowest ratio, we must eliminate meters from each of the other bins until all bins have a 1.0 ratio between treatment and comparison groups. This can be done in one step by dividing the comparison pool meter count in the limiting bin by the treatment group fraction of meters in the limiting bin and then multiplying by the fraction of meters in the treatment group for each bin. Using the lower-left bin as an example:

$$Comparison\ Group\ Meter\ Count_{Lower\ Left} = 0.29 \times \frac{17}{0.21} = 23$$

Applying this procedure to all bins yields the following comparison group:



We see that in this case we have produced a proportional match in each bin by resampling a comparison pool of 190 meters to a comparison group of 79 meters.

It is important to note that achieving proportional binning does not equate to an unbiased sample, even when only considering the specific stratification parameters. In this example, it is possible that within the top bins the comparison pool is over-represented from 1000 to 3000 annual kWh and under-represented from 3000 to 5000 annual kWh. In this case, it may be necessary to move to three or more bins for annual kWh to force a more granular sampling.

This is a key concept we will return to below. Stratified sampling can be made more precise with enhanced granularity - increased number of parameters and/or an increased number of bins per parameter. However, as with many other sampling techniques, stratified sampling is ultimately constrained by the total number of meters in the comparison pool and the number of meters needed for a comparison group. With a smaller comparison pool, fewer bins across fewer parameters can be utilized before eliminating too many meters to meet the minimum requirement for the final comparison group.

B. Enhanced Stratified Sampling

How does one define the success or failure of a comparison group formed by stratified sampling or any other method for that matter? Is a comparison group reliable and representative *because* we achieve

proportional binning for certain parameters in relation to a treatment group? What is the best or proper way to gauge statistical similarity between a comparison group and a treatment group?

Stratified sampling is a means to an end. The ultimate goal is a comparison group that is the most representative of a treatment group and with enough statistical power as to not introduce undue uncertainty into a savings measurement. To this end, stratified sampling is just a tool. A good proportional match for chosen stratification parameters does not guarantee that other important aspects of energy usage have been well emulated by the selected comparison group.

Many evaluators, including Recurve, have utilized individual site-based matching schemes in which each treatment meter is assigned one or more comparison group meters based on a direct measurement of usage. Site-based matching has been done in recent studies by minimizing a Euclidean distance metric computed across the baseline period monthly consumption for each comparison group meter tested against each treatment group meter. In site-based matching the fact that this strategy focuses directly on the load profiles of treatment and comparison meters - and not isolated usage characteristics - is a highly attractive feature.

With these considerations in mind, the last selection method we cover in detail here is advanced stratified sampling wherein optimization is not conducted by minimizing treatment vs. comparison discrepancies specific to the chosen stratification parameters themselves, but rather by optimizing the fit between the resultant load profiles. Therefore, instead of striving for statistical equivalence using a T-Test or KS-Test or enforcing an optimization scheme on the stratification parameter distributions, we propose gauging the performance of a candidate stratified sampling scheme on its ability to minimize the discrepancy between treatment and comparison group load profiles.

Where only monthly data are available, the load profile has a maximum of 12 data points per customer per year. In contrast, with hourly data one could attempt to optimize across the entire 8,760 annual load shape, but this would be enormously expensive and not likely to yield significantly improved results compared to more aggregated options. With hourly data, we suggest that taking into account both seasonal and weekday vs. weekend differences is important beyond simply assessing an average 24-hour average daily load shape. Thus while 8,760 data points may be excessive, 24 cannot capture important comparative features. Instead, we recommend using average seasonal weekly load shapes for the summer, winter, and shoulder timeframes. The resulting 504 data point representation is nearly 20 times smaller than the full 8,760 profile yet retains most of the relevant information.

In taking this approach it is important to guard against the potential pitfall that wildly disparate comparison group load shapes could simply average to produce a similar profile to an average treatment group load shape. Therefore, a straight least-squares optimization of an average comparison group load profile vs. an average treatment group load profile should be avoided. Instead, we recommend an approach in which each data point in the average load profile is broken into bins of equal proportion for both treatment and comparison groups. Considering a monthly load profile, the average January consumption for treatment and comparison group customers is split into bins by percentile. A treatment group of 250 meters can be ordered by January consumption and broken into 10 groups of 25 meters.

The lowest usage group corresponds to the 0 - 10% decile and the highest usage group corresponds to the 90 - 100% decile. The exact same procedure for the comparison group yields a corresponding set of deciles. The average January usage for each decile in the treatment group forms a distribution that can be compared directly to the distribution produced from the comparison group. At this point a sum of squares computation can be performed across each decile. Finer binning can also be conducted if computational resources allow.

Continuing with the monthly example, repeating this process for each month yields 12 sums of squares that can be summed for a total sum of squares value, which represents the degree of distributional similarity between treatment and comparison groups. As described above, when hourly data are available, seasonal load shapes can be the focus of this computation. With this approach we can ensure that an average comparison group load profile does not appear to be a good representation of a treatment group when the underlying usage distributions among component meters differ substantially.

In the next section we turn our attention to the automated development of candidate stratification schemes to be tested against a treatment group using the approach described here. The stratification scheme that produces the lowest value of the summed least-squares computation will be used to produce the comparison group.

C. Automating Generation of Candidate Stratification Schemes

In the illustrative example of Section A several areas of potential subjectivity are apparent:

1. The number of parameters
2. The choice of specific parameters
3. The number of bins
4. Where to place the bin limits
5. The number of comparison group meters to select

As many aspects of this approach should be standardized and automated as possible in order for these methods to be consistently and routinely applied. For Version 1.0 of these comparison group methods we make recommendations and provide open-source code to fully automate the first, third, and fourth of these three factors and we have provided analysis to inform recommendations for the last point. We also provide recommendations for the second point but must leave a fully automated framework to address this question for possible refinements and a future updated version of both methods and code.

1. The number of parameters

In considering the number of parameters it is important to understand why selecting many parameters is infeasible. In Section A we introduced the concept of a “limiting bin,” where the comparison pool is most underrepresented relative to a treatment group. As more parameters are added, the number of bins grows exponentially. Imagine adding parameters, each with only 2 bins. Each new parameter interacts with all other parameters, thereby doubling the number of total bins. Therefore, going from 2

to 3 parameters increases the number of bins from 4 to 8. But increasing from 6 to 7 parameters increases the number of bins from 64 to 128. As the number of bins increases, the probability that the limiting bin severely restricts the possible number of comparison meters increases as the comparison group is carved into finer and finer slices.

For most stratification schemes we expect that a maximum of three parameters can provide for sampling over several important aspects of customer usage while avoiding rapid over-binning. However, where sufficient data are available, there is no harm necessarily in moving to more parameters.

2. The choice of specific parameters

While at this point we do not offer concrete parameters for all use cases or code to automate parameter selection, we do provide the following considerations and guidance:

- Parameters should be chosen that are relevant to the program because they are likely to be sensitive to the specific intervention. For instance, if a program plans to replace inefficient gas furnaces, then choosing parameters related to space heating gas usage, such as temperature-dependent or winter gas consumption, would help ensure that the comparison group reflects the usage characteristics most likely to showcase program influence.
- Parameters should be chosen that are not themselves highly correlated. For MCE's Residential customer base, we have measured the correlation coefficient between summer and annual electricity consumption to be 0.94. Thus using both of these metrics as stratification parameters would offer very little additional information.
- A combination of dimensioned and dimensionless metrics should be used. A dimensioned parameter will reflect total consumption while a dimensionless parameter will allow a focus on critical aspects of consumption that are more related to *how* customers are using energy instead of just serving as another gauge of how much they are consuming. This recommendation is related to the last point. The correlation coefficient between annual kWh and cooling kWh is 0.56 but is only 0.15 between annual kWh and percent cooling.[20] If a program is focused on air conditioning, the latter combination of parameters would provide for a higher level of distinction.

3, 4, 5. The number of bins, bin limits, and the number of comparison group meters

These items are interrelated and we cover them together. We have automated an optimized binning scheme with the following procedure:

i. For each parameter, the minimum of the lowest bin is set by the minimum value observed in the treatment group. Similarly, the maximum value of the highest bin is set by the maximum value observed in the treatment group.

ii. A minimum of 1 and maximum of 8 bins are allowed per stratification parameter.

[20] The percentage of a customer's total annual usage that is found to be temperature-dependent with warm weather.

iii. Beginning with a single bin for each parameter, every possible binning combination is scanned. For two parameters there are 64 possible binning arrangements [(1, 1), (1, 2), (2, 1), (2, 2)...(8, 8)].

iv. Depending on the size of the treatment group, scans are aborted for binning combinations that fail to yield a user-defined ratio of comparison group meters to treatment group meters. A minimum ratio of 4:1 is recommended for small Residential treatment groups (< 750 meters).[21] For the Commercial sector this ratio will depend on the business type.

v. For binning combinations that yield a large number of meters, a random selection of available comparison pool meters is taken to meet a user-defined maximum. For both Residential and Commercial programs, a maximum value of at least 3,000 meters can help ensure uncertainty due to random variability in the comparison group is kept under +/- 2% in 90% of cases (see Fig. 7 of Chapter 2).

vi. All candidate comparison groups that pass the minimum threshold of step iv are passed to the sum of squares calculation described in Section IV.B.

vii. The final comparison group is selected based on the lowest sum of squares value computed as described in Section IV.B.

With the approaches described in this chapter, the establishment of comparison groups can be use-case specific and completed on the basis of a forecasted or actual treatment group. In the commercial sector the most important factor in achieving a comparison group capable of isolating program impacts and removing COVID impacts should focus first on building type wherever those data are available. In the Residential sector, geographic location and key usage patterns can serve as the basis for formulating comparison groups capable of reducing COVID-related residuals in a difference of differences calculation.

In both the Commercial and Residential sectors, where sampling stems from a program's actual participant group, the enhanced stratified sampling methods of sections B and C are designed to strike a balance between the computational feasibility of stratified sampling and the advantages of direct load profile matching offered by site-based strategies. Where hourly data are available, an optimization conducted on seasonal weekly load shapes with largely independent stratification parameters that are representative of differences between a treatment group and comparison pool promises to produce comparison groups that can be used with confidence.

D. Example

As a test case for these methods we created a fake treatment group, which differed from the general population of MCE residential customers, and then executed each of the above steps to automatically

[21] Assuming Poisson statistics a 4:1 ratio will ensure the comparison group contributes no more than approximately a third of the uncertainty in the savings calculation.

select a comparison group. The treatment group was pulled from the first 50,000-meter sample described in the previous chapter with the following steps:

2. Customers were selected who were in the top 75% of total baseline period usage and the top 40% of their utility cost per MWh ratio. The latter metric Recurve calculates based on customers' avoided cost profiles using the California Public Utility Commission's 2020 avoided cost data.[22] Out of the initial 50,000-meter sample, 16,606 meters met these thresholds.
3. Of these 16,606 meters a random sample of 2,000 meters was selected.
4. The second 50,000-meter sample (see Chapter 2) was utilized as the comparison pool.
5. Because customers with higher utility cost per MWh tend to use more during the summer peak period and a visual inspection of the seasonal load shapes indicated these customers also had a steeper evening ramp than an average customer, the three stratification parameters chosen were annual kWh, percentage of kWh during summer peak[23] and evening ramp ratio.[24]
6. For this exercise, maximum bins for each stratification parameter were set to 3, though this limit could be made significantly higher if beneficial.
7. A target of 5,000 comparison group meters was set.

The binning scheme that yielded the lowest sum of squares across the 504 seasonal weekly load shape profile was 2 bins for annual kWh, 3 bins for percent summer peak, and 2 bins for the evening ramp ratio with a resulting 4,997 meters. This scheme improved the comparison pool sum of squares metric from 561 to 66 for the final comparison group.

The left-hand plot of Figure 15 shows the impact of stratification along each parameter that results from this three-dimensional binning scheme. The distribution of the comparison pool, shown in red, is significantly different, especially for the percent summer peak parameter, than those of the treatment group. After stratified sampling, clear improvements are observed across the board, though there is still some mismatch apparent in the lowest percent summer peak bin. This would likely be remedied with a higher limit on bins for this parameter.

[22] https://www.cpuc.ca.gov/General.aspx?id=5267

[23] Defined as the percentage of a customer's total annual kWh usage that occurs from 4 - 9 pm during the months of June - September.

[24] Defined as a customer's average usage during hour 18 minus average usage during hour 14 divided by the total annual usage.

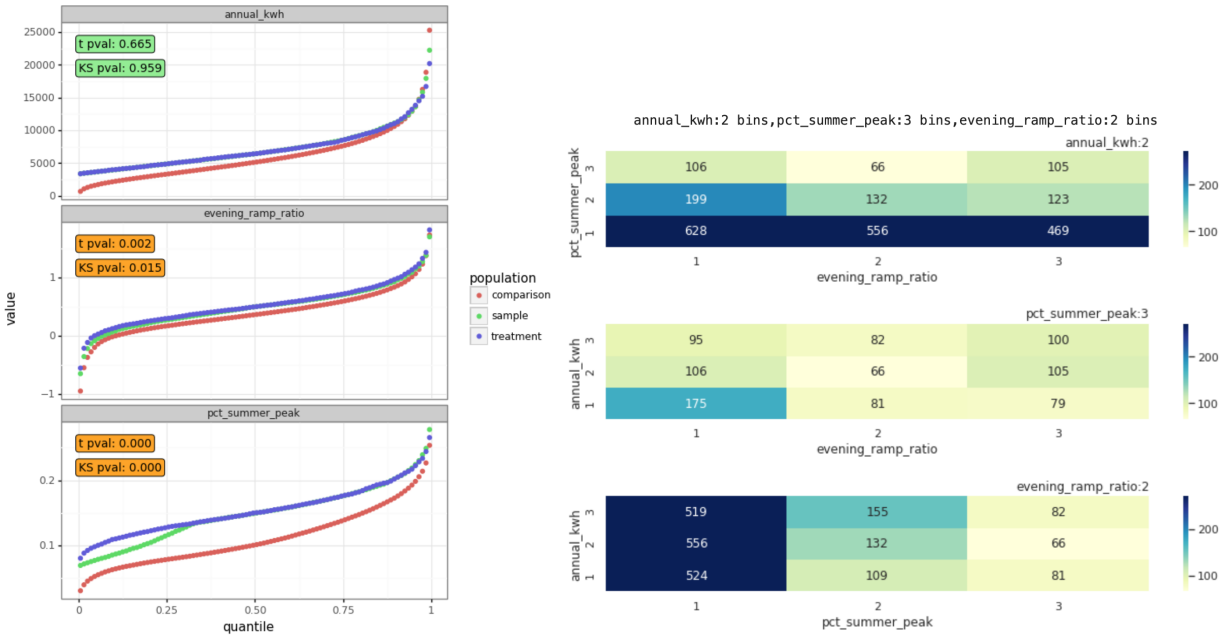


Figure 15: Left**:** Quantile plots showing the impact of stratification along each parameter that results from the chosen three-dimensional binning scheme Right: Heatmaps that show the sum of squares metric for some of the binning combinations that were searched.

The right-hand plot of Figure 15 are heatmaps that show the sum of squares metric for some of the binning combinations that were searched. Steady improvement can generally be seen from left to right and from bottom to top as the binning is made finer.

Figure 16 shows the full seasonal weekly load shapes for the treatment group, comparison pool, and comparison group. Clear improvement is seen as a result of the stratification and optimization steps.

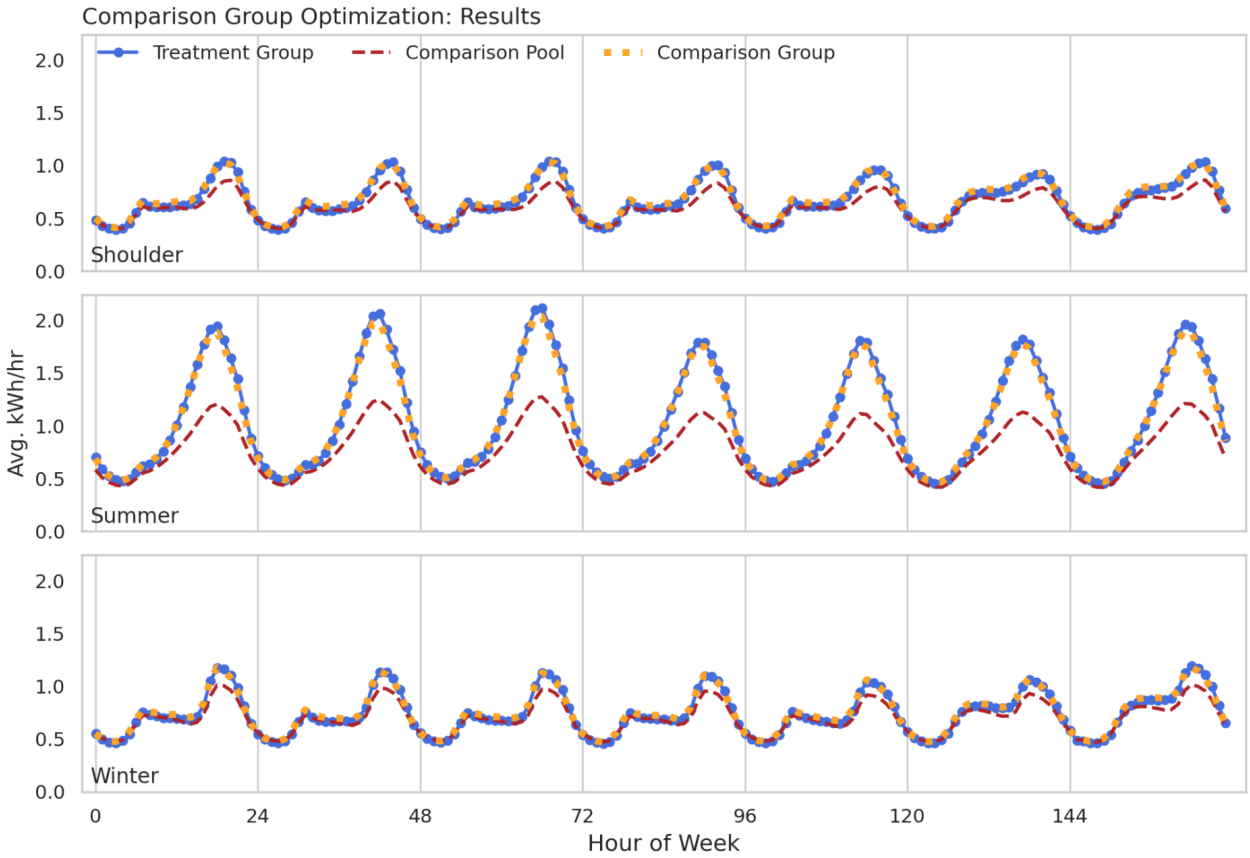


Figure 16: Baseline seasonal weekly load shapes for the treatment group, comparison pool, and comparison group resulting from the advanced stratified sampling scheme described in this chapter.

Just as important a test is how the comparison group behaves during the counterfactual period relative to the treatment group. Figure 17 shows that the good load shape match observed in the baseline period continues into the reporting period (the COVID period).

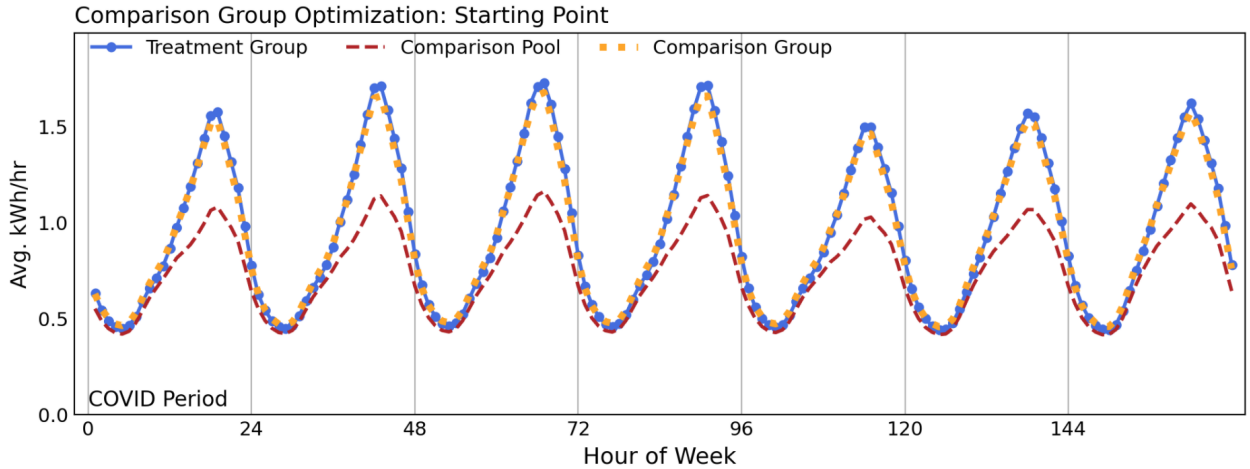


Figure 17: Reporting period (COVID-period) weekly load shapes for the treatment group, comparison pool, and comparison group resulting from the advanced stratified sampling scheme described in this chapter.

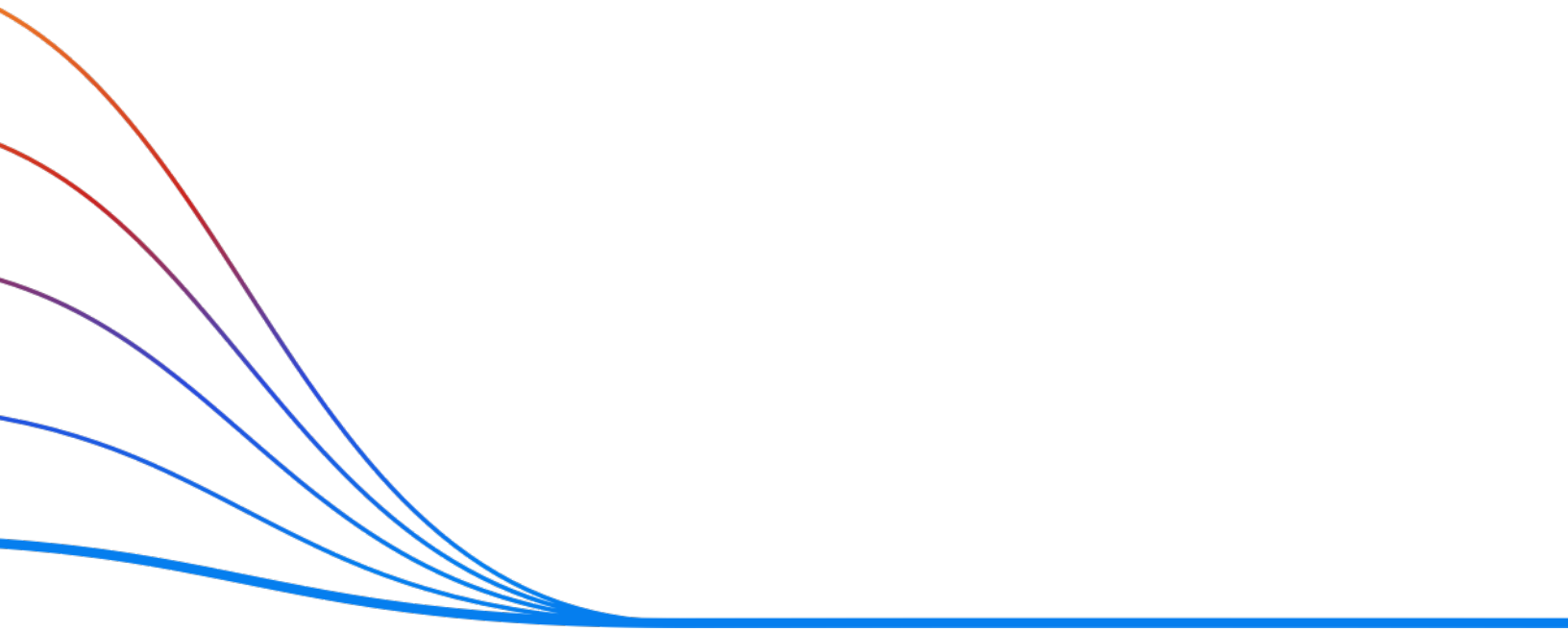
VI. Open Source GRIDmeter Codebase

As part of this research effort, Recurve has compiled an open-source codebase (GRIDmeter), which allows users to execute the sampling methods described here. The GRIDmeter codebase is available to all parties.[25]

[25] https://grid.recurve.com/

RECURVE
SHAPE THE FUTURE OF ENERGY

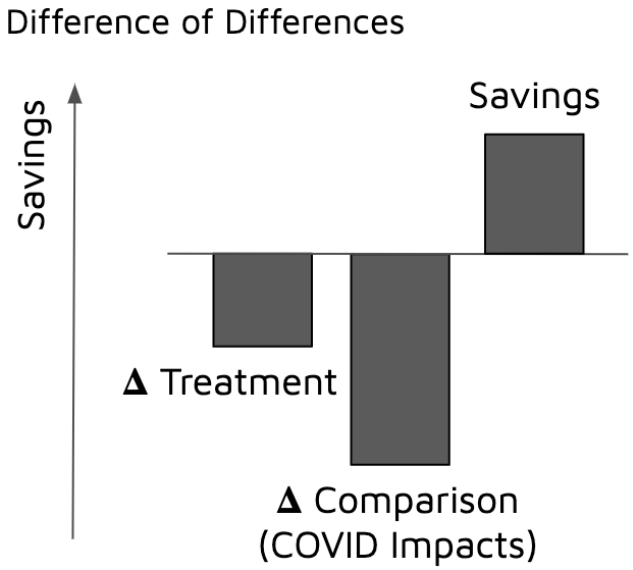
Chapter 4: Load Impact Calculations via Difference of Differences



I. Introduction

When using a comparison group, the final savings calculation is often referred to as a "difference of differences." The concept of a difference of differences is relatively straightforward, but if not specified in sufficient detail many different implementations - and answers - are possible. This chapter describes the difference of differences calculation and specifies important implementation details.

The following schematic illustrates the core concept of a difference of differences calculation:



This schematic shows a case in which participant energy usage increased after program intervention. With a comparison group in hand, the difference of differences calculation consists of three basic steps:

1. *Measure the change in consumption for program participants.* This step is identical to a savings calculation absent a comparison group. In this step a baseline period model is developed for treatment group meters. This model is projected into the reporting period as the counterfactual. The counterfactual is the prediction of energy consumption that would have existed without the program. Subtracting the counterfactual from the actual post-program usage in the treatment group yields a measurement of savings and yields the first "difference" in the difference of differences calculation:

$$Diff_{Treatment} = Counterfactual_{Treatment} - Observed_{Treatment}$$

2. *Measure the change in consumption for the comparison group.* This step is analogous to step 1 with the measurement conducted on the comparison group. The $Difference_{Comparison}$ represents the exogenous trends in the broader population and, with a well-designed comparison group, will capture the impacts from COVID and other exogenous factors.

$$Diff_{Comparison} = Counterfactual_{Comparison} - Observed_{Comparison}$$

3. *Compute savings*. With the change in consumption figured for both treatment and comparison groups, the program's impacts are then calculated by adjusting the treatment group savings for

the naturally occurring savings observed in the comparison group.

$$Savings_{(Diff of\ Diffs)} = Diff_{Treatment} - Diff_{Comparison}$$
$$= (Counterfactual_{Treatment} - Obs_{Treatment}) - (Counterfactual_{Comparison} - Obs_{Comparison})$$

In the schematic above, the treatment group experienced *increased* usage after program participation. However, the comparison group exhibited an even greater consumption increase, indicating that the program produced positive savings. For residential programs in-field when COVID impacts were peaked, this is a very realistic scenario. As described in Chapter 1, Recurve has observed the average residential customer in MCE territory has increased electricity usage by 7.9% on account of COVID. Recurve has observed similar results in the assessment of gas usage for other program administrators. For a program saving 7%, a savings calculation without a comparison group over this same timeframe would yield -1.2% savings.

Figure 18 shows each element of a difference of differences calculation for hypothetical treatment and comparison groups. The curves in the top panel show the observed and counterfactual weekly load shapes of a treatment meter. The difference between the two is calculated and shown as the gray trace. The middle panel gives the same information for the comparison group. Finally, the bottom trace gives the average savings (difference of differences) by hour of week.

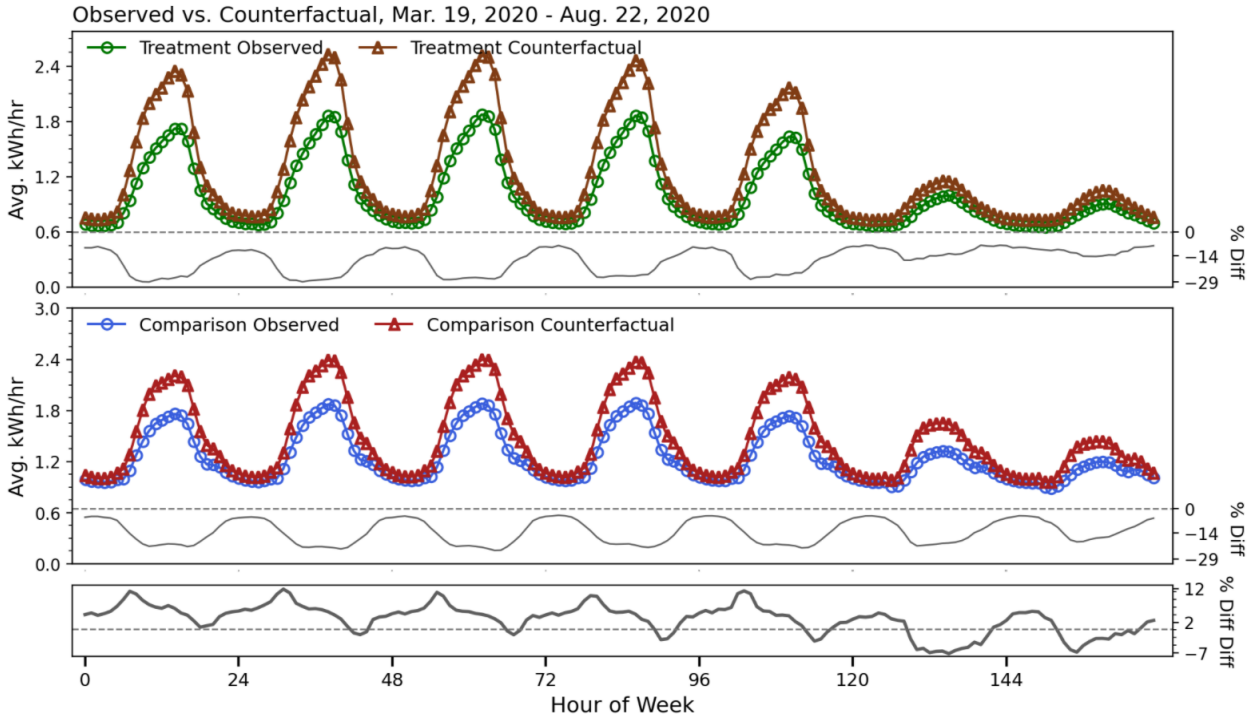


Figure 18: Average weekly load shapes and percent difference traces for each element of a difference of differences calculation for hypothetical treatment and comparison groups

In this example we see that the counterfactual for both the treatment and comparison groups was significantly higher than the observed consumption. This type of scenario is expected to be common for Commercial programs in-field today. Recurve has measured a 15.0% decline in Commercial sector electricity consumption during the COVID period. (See Chapter 1 for more details.)

II. Program Considerations and Implementation

While the core concepts of a difference of differences calculation are straightforward for those well versed in comparison group theory, in practice several practical questions emerge, including:

- Should the difference of differences calculation be conducted on an absolute or percentage basis and why?
- If done on a percentage basis, what value should the resulting percent savings be multiplied by to determine absolute savings?
- If done on a percentage basis, how should the equation be formulated to handle customers without distributed generation, in which every value is greater than or equal to 0, and customers with distributed generation in which observed and/or modeled values can be negative?
- How should one account for the staggered project installation dates of a real-world program in determining baseline and "reporting" period dates for the comparison group?

We start with the first of these questions.

A. To mitigate risk and allow for more flexible comparison group selection, the difference of differences calculation should be conducted on a percentage basis.

Consider the following table, which provides a hypothetical example of a savings calculation for a treatment and comparison group:

	Treatment Group	Comparison Group
Avg. Annual MWh Baseline	100	150
Avg. Annual MWh Reporting	90	138
Difference	10	12
Savings	**-2**	
% Difference	10%	8%
Savings	**2**	

The average treatment group customer used 100 MWh in the baseline period and 90 MWh in the reporting period for a difference of 10 MWh. The average comparison group customer used 150 MWh in the baseline period and 138 MWh in the reporting period for a difference of 12 MWh. If taking the absolute difference of differences we would find that the treatment group customer had negative savings (-2 MWh). However, on a percentage basis, the average treatment group customer used 10% less while the average comparison group customer used 8% less. If savings are calculated via these percentages, we find that a 2% positive savings value should be ascribed to the program.

Importantly, the average comparison group customer was larger than the average treatment group customer. This led to a smaller percentage change in usage producing a larger total change in

consumption. While it may be true that the comparison group in this case is clearly not a perfect representation of the treatment group, a program should not be so directly penalized for such a mismatch. With a savings calculation instead conducted on a percentage basis, error from a skewed comparison group is contained to a second-order effect.

B. With a percent difference of differences calculation, final savings should be determined via multiplying by the treatment group counterfactual.

While it is important to mitigate risk in the difference of differences calculation by performing the computation on a percentage basis, there is no obvious or perfect answer to the question of what that percentage should ultimately be multiplied by to produce a final savings value. If multiplying by the reporting period observed consumption, the program is penalized for the very savings it produces. If multiplying by the counterfactual, COVID impacts are essentially ignored despite the fact they are obviously real. Multiplying by baseline period usage or the baseline model has the same pitfall. One could envision a hybrid approach in which the percent difference of differences is applied to the combination of treatment group counterfactual adjusted for the COVID impacts observed in the comparison group. However, this level of abstraction introduces unnecessary complexity and does nothing to forward the goal of enabling certainty needed to design and implement meter-based programs.

Here we recommend using the treatment group counterfactual to compute absolute savings from the percent difference of differences calculation. We make this recommendation for two primary reasons:

1. This would be the most sensible and justifiable approach in the absence of a large exogenous event like COVID. As time goes on, COVID impacts should diminish (we already see evidence that COVID impacts have abated over the last several months in MCE data and elsewhere).

2. Most program interventions produce savings with expected persistence of several years or more. The “lifecycle” savings that result from a meter-based measurement are often determined by applying the first-year savings calculation across the expected lifetime of the measure. For longer-lived program impacts using the treatment group counterfactual can help ensure the first-year savings measurement is most appropriate for application to a lifecycle savings calculation, despite COVID.

C. Absolute % Difference of Differences

If all observed and model values are positive, then a percent Difference of Differences (%DoD) formulation is straightforward. In this case %DoD accounts for mismatches between the treatment and comparison group load shapes by first computing the difference between the fractional savings of the treatment group and comparison group multiplied by the treatment group counterfactual. This step acts to “correct” the treatment meter counterfactual for the model error patterns observed within the comparison group during the reporting period. Savings are then calculated as the difference between this corrected counterfactual and the observed treatment consumption. These steps are combined in

Eq. 10 where s is savings, o_{CG} and m_{CG} are the comparison group observed and counterfactual values, and o_T and m_T are the treatment group observed and counterfactual values.

$$s = \frac{o_{CG}}{m_{CG}} m_T - o_T \quad (10)$$

That is, the corrected savings is computed by rescaling the counterfactual with a correction factor that is equal to the comparison group's observed-counterfactual ratio. This formulation leads to comparison group corrections to savings (c) given by Eq. 11.

$$c = m_T(\frac{o_{CG}}{m_{CG}} - 1) \quad (11)$$

However, the savings correction fails in both direction and magnitude given various combinations of signs between o_{CG}, m_{CG}, and m_T. For instance, if the comparison group observed (o_{CG}) and counterfactual (m_{CG}) values are positive, but if the treatment group counterfactual (m_T) is negative then the %DoD correction is in the wrong direction and with improper scale.

Solution: Absolute Percent Difference in Differences

The directionality and magnitude issue can be addressed by introducing absolute values as in Eq. 12.

$$s = -|m_T| \frac{m_{CG} - o_{CG}}{|m_{CG}|} + m_T - o_T \quad (12)$$

We call this approach "absolute percent difference on differences" (abs%DoD). In effect, we are rescaling the comparison group "savings" by the ratio of the *absolute values* of the counterfactuals. In cases where o_{CG}, m_{CG}, and m_T are all of the same sign, Eq. 12 simplifies to Eq. 10.

The abs%DoD approach is mathematically consistent given meter data that combines distributed generation and full building consumption. This approach eliminates sign flip issues and the magnitude of the comparison group correction can still produce scaling issues when analyzing distributed generation customers, especially when counterfactual values are close to 0. When m_{CG} is close to zero, the denominator of Eq. 12 can blow up to arbitrarily large values, creating unreasonable spikes in the corrected savings. We recommend implementing a cap on the correction factor to protect against such cases. When m_T is close to 0, the comparison group correction will be small regardless of the comparison group behavior. These topics should be researched further. It is our assessment that the only fully valid solution in all cases is achievable through disaggregation of distributed generation and calculations based only on the reconstituted whole building load.

D. Baseline and reporting period comparison group calculations should closely mirror the range of treatment group intervention dates.

Energy usage patterns change over time due to economic conditions, changing technologies, population dynamics, and global pandemics, among other factors. The very purpose of a comparison group is to provide a measurement of these exogenous factors that can be immediately applied to best isolate program impacts. For this reason, it is important to align the timeline of comparison group calculations

to the dates of a program's participation. As a practical matter this means it is not sufficient to select a comparison group and then simply compute savings for this group for one set of baseline and reporting period dates while the program subject to comparison group adjustment served customers at various points throughout the year. On the other hand, selecting an entirely different comparison group for each day or each week may be too expensive and impractical. Therefore, we recommend that at a minimum computing the savings of a single comparison group should be done monthly to best capture the appropriate timelines for a program. For more discussion on this point see section II.D of Chapter 3.

E. A Note on Computational Granularity

Because a percentage difference of differences calculation is a nonlinear adjustment to savings, the granularity of summation may produce varying results. This means that calculating savings using different temporal comparison group adjustments may produce different savings when aggregated. As a best practice we recommend calculating savings on the most granular temporal level possible. For example, if hourly data are available, savings should be calculated on an hourly percentage basis, and then summed, rather than daily or monthly.

This principle likely applies within comparison groups as well. If a stratified sampling method was used to produce the comparison group, the same stratification should be used to calculate the percent savings prior to summation. For site-based matching, savings for each site can be calculated using the individual matches and then summed to produce final results.

F. Absolute % Difference of Differences Using Clustering

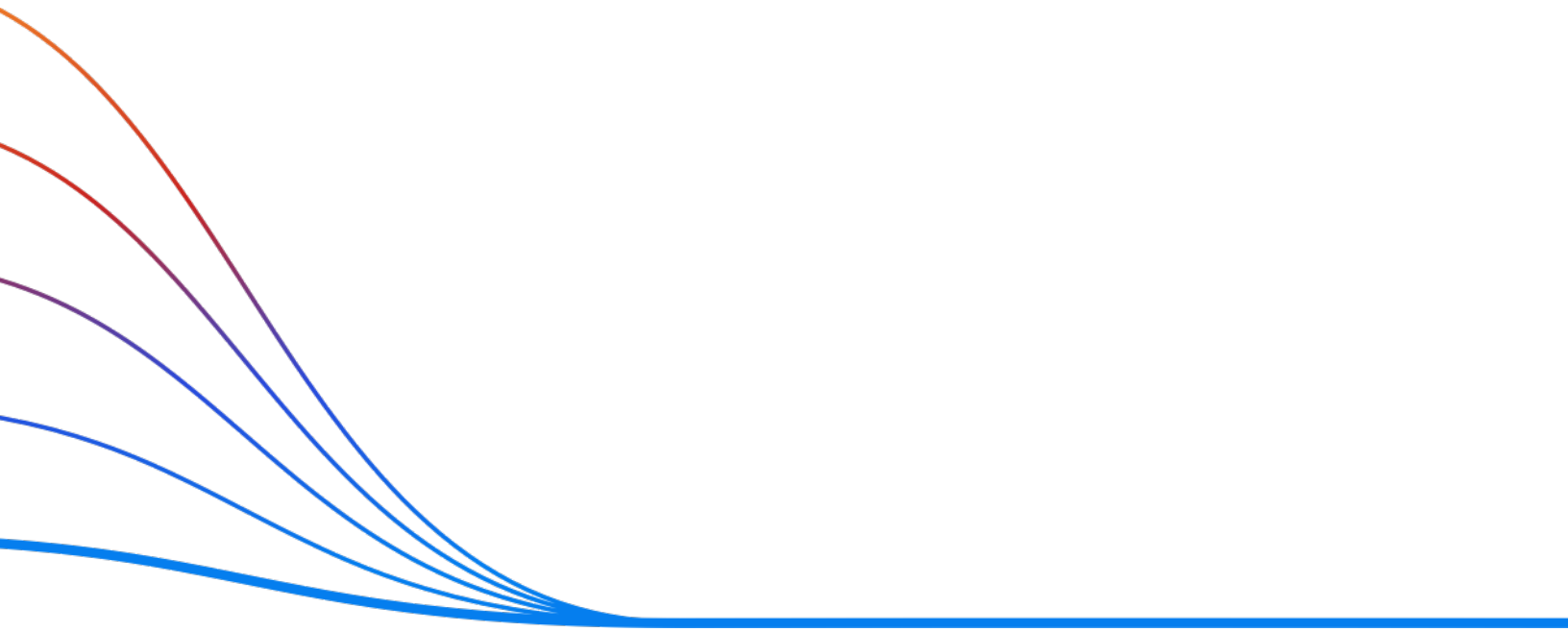
In Clustering the comparison group for each treatment meter is assigned as a weighted combination of clusters. Each cluster will consist of a different number of meters. This raises the need to specify further how the comparison group corrections should be calculated. In short, the comparison group correction should be influenced by a given cluster to a degree that corresponds to that cluster's weight.

In a trivial example, if a treatment meter is assigned to two clusters with weights of 0.7 and 0.3, the final correction should be taken as the sum of 70% and 30% of each cluster's correction respectively.

In addition, with the site-level corrections of clustering, the implicit weighing by consumption is no longer desirable because each treatment meter instead has its own unique correction factor for each hour. Instead, correction factors are calculated as $mean\left(\frac{o_{CG}}{m_{CG}}\right)$. The change to averaging the ratio might seem small but is quite impactful and allows the relationship between the o_{CG} and m_{CG} to be fully utilized.

RECURVE
SHAPE THE FUTURE OF ENERGY

Chapter 5: Quantifying Residuals and Variance in the Residential Sector



I. Summary

This chapter details residual measurements from difference of differences calculations for MCE's residential sector. We describe the difference in consumption between forecasted and observed as a residual, which can be understood as the combination of exogenous factors and statistical noise after weather-normalizing the data. Note that a value of percentage residual is in reference to total consumption. Unintended residuals between program participants and potential comparison groups could arise due to different geographic locations, different usage patterns and other factors such as income and demographic characteristics. In this chapter we focus on geography and usage patterns with the goal of understanding to what extent misalignments between treatment and comparison groups would be expected to yield savings uncertainty and variance on account of COVID.

A. Geography

This section summarizes results for the geographic trials. Random samples were taken from each of the six largest cities in MCE service territory. These cities cover a diverse range of climates as well as income and demographic characteristics. For instance, the city of Richmond has nearly double the proportion of low-income residents than Napa and has far lower average usage than in MCE territory as a whole.[26] For each of these samples we assess residuals in the difference of differences calculation when the following strategies are employed for comparison group selection:

1. No comparison group
2. A randomly selected comparison group of residential customers from across MCE territory
3. A randomly selected comparison group of non-overlapping customers from the same city.

Figure 19 gives an example of the load shape differences observed between one of these cities. The average daily load shape of an MCE residential meter is shown in blue (circles) with the average daily load shape of a residential Pittsburgh meter shown in green (triangles).

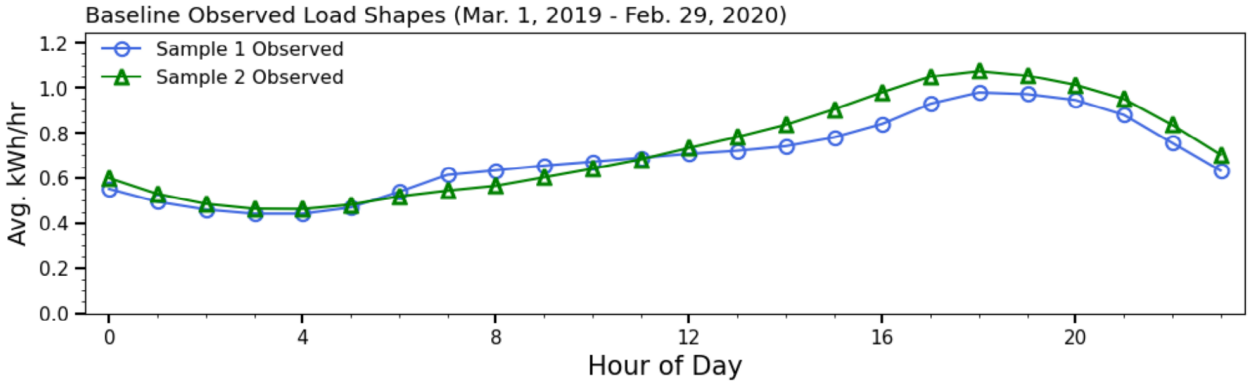


Figure 19: Average daily load shape of a residential non-solar MCE meter (blue circles) and a residential non-solar MCE meter in Pittsburgh (green triangles).

[26] 23% of Napa's residential non-solar meters are associated with customers enrolled in the California Alternate Rates for Energy Program (CARE) rates compared to 38% for Richmond.

The difference in load shape may or may not lead to a COVID-related residual in % difference of differences calculation. Summary results from this experiment are provided in Table 4 for both residuals present in a total savings calculation, and the mean absolute percentage error (MAPE) observed in the measurement of hourly load impacts.

Table 4

COVID Period		City* vs Random			City vs City^		
City	Sample	% Diff	% Diff Diff	Hourly MAPE (%)	% Diff	% Diff Diff	Hourly MAPE (%)
Concord	Random	8.15	-2.26	4.09	10.11	1.08	1.89
	City	10.41			9.03		
Napa	Random	8.15	3.79	6.23	5.64	0.01	2.18
	City	4.36			5.63		
Pittsburgh	Random	8.15	-1.76	5.88	10.63	0.88	2.45
	City	9.90			9.75		
Richmond	Random	8.15	-0.38	6.18	8.38	0.43	2.09
	City	8.53			7.95		
San Ramon	Random	8.15	-0.55	5.65	10.31	0.79	2.43
	City	8.70			9.52		
Walnut Creek	Random	8.15	0.05	3.58	8.80	0.12	2.25
	City	8.10			8.68		

*Random sample size = 50,000 meters, City sample size = 3,000 meters except for Pittsburgh (2,500 meters)
^City sample 1 = 3,000 meters except for Pittsburgh (2,500 meters)
City sample 2: (Concord = 5,124, Napa = 3,433, Pittsburgh = 2,468, Richmond = 4,049, San Ramon = 3,170, Walnut Creek = 3,600)

Without a comparison group, residuals in a total savings calculation ranged from -5.6% to 10.6% across different cities. When comparing a random sample from MCE's entire service territory, residuals ranged from near 0 (Walnut Creek) to 3.8% (Napa). This degree of uncertainty may be acceptable to program administrators for whom measuring annual savings is the most important consideration. Despite a smaller sample size, a reduction of residuals is observed in most cases when a comparison group is formulated by sampling from the same city. In all cases investigated here, the residual in the COVID-period total difference of differences calculation was less than 1.1% when selecting treatment and comparison randomly from the same city.

For program administrators seeking reliability in the hourly calculation of load impacts, these results show a clear advantage of pulling the comparison group from the same geographic location. Mean Absolute Percent Error (MAPE) in the hourly difference of differences measurements was below 2.5% for each within-city trial but ranged from 3.6% to 6.2% when comparing a specific city to the territory-wide sample. When moving from a sector-wide to a city-specific comparison group, the improvement in hourly measurements is evident in the data provided in Appendix B.

B. Usage Characteristics

Along with geographic considerations, demand-side programs often target customers based on specific usage patterns. For example, a demand response program would likely seek customers who exhibit high peak period usage. Customers with different usage patterns may respond differently to COVID and if not accounted for these differences can lead to bias in a difference of differences calculation.

In this section we establish samples of MCE residential customers with systematic differences in particular usage characteristics, measured during the pre-COVID-period. For each sample, we then test the following comparison group scenarios:

1. No comparison group
2. A randomly selected comparison group of residential customers from across MCE territory
3. A randomly selected comparison group of customers who meet the same consumption-based selection criteria.

Table 5 details the selection schemes explored here.

Table 5

Sample	Parameter 1	Threshold 1*	Parameter 2	Threshold 2*
1	annual_kwh	>= 0.25	pct_cooling	>= 0.6
2	annual_kwh	>= 0.25	pct_summer_peak	>= 0.6
3	pct_baseload	>= 0.75		
4	pct_discretionary	>= 0.75		
5	evening_ramp	<= 0.6	evening_ramp_ratio	<= 0.4
6	shldr_midday_restofday_ratio	>= 0.6	pct_winter_morn	>= 0.6

*These thresholds correspond to percentiles. For example a meter is eligible for the first sample if it is in the top 75% of annual kWh and top 40% of the percentage of usage from cooling among all MCE non-solar residential customers.

Figure 20 gives an example of the load shape differences observed between an average MCE customer and a customer in Sample 1 (Table 3). The average daily load shape of an MCE residential customer is shown in blue (circles) with the average daily load shape of a customer in Sample 1 in green (triangles).

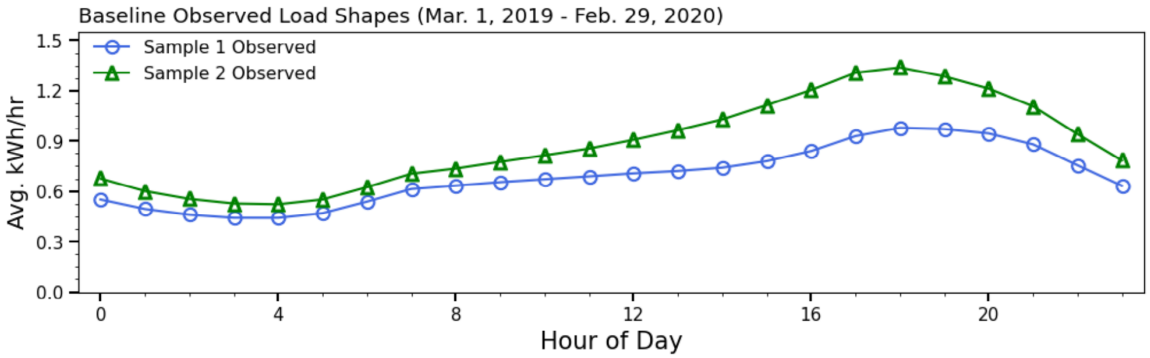


Figure 20: Average load shape of a residential non-solar MCE meter (blue circles) and a residential non-solar MCE meter in the top 75% and top 40% of all residential meters in annual usage and the percent of usage from cooling.

Table 6 provides a summary of results for these comparison group tests.

Table 6

COVID Period

	Selected* vs Random			Selected vs Selected^		
Sample	% Diff	% Diff Diff	Hourly MAPE (%)	% Diff	% Diff Diff	Hourly MAPE (%)
Random	8.15	0.80	3.32	7.30	-0.08	1.47
Sample 1	7.35			7.37		
Random	8.15	0.54	3.52	7.77	0.01	1.32
Sample 2	7.61			7.76		
Random	8.15	1.48	2.85	6.63	0.34	1.28
Sample 3	6.67			6.29		
Random	8.15	2.00	4.05	6.28	0.62	1.91
Sample 4	6.14			5.65		
Random	8.15	2.61	3.93	6.17	-0.11	1.54
Sample 5	5.53			6.28		
Random	8.15	1.53	3.36	6.19	0.50	1.55
Sample 6	6.62			5.69		

See Table 1 for details. The Random sample was 50,000 meters. The first "Selected" sample was 3,000 meters
^The second "Selected" samples had the following meter counts (16,718, 17,135, 12,500, 12,500, 18,787, 9,577)

Without a comparison group, residuals ranged from -5.7% to 7.8% across the different samples of Table 6. Interestingly, none of these samples exhibited greater impacts from COVID than the territory-wide random selection (8.2%). When using this territory-wide random sample as a comparison group, residuals in the difference of differences calculation ranged from 0.5% to 2.6%. Reminiscent of the geographic samples, despite smaller sample sizes, a reduction in residuals is observed in all cases when a comparison group is formulated by sampling with the same selection criteria. In the current cases, residuals in the COVID-period total difference of differences calculation were less than 0.6% across the board when doing so.

Significant improvements in the hourly MAPE are observed when employing the same usage-based selection criteria between samples. With the random comparison group approach the MAPE ranged from 2.9% to 4.1% compared to 1.3% to 1.9% when applying the same selection requirements.

II. Experimental Details

The following stepwise analysis was conducted to produce the results above and gauge the degree of residual in a % difference of differences calculation due to COVID.

1. Hourly OpenEEmeter 2.0 calculations were performed on all residential meters in MCE territory using the timeline of Figure 1.

Meters were eliminated for which any of the following criteria were true:

   a. Were solar customers (identified by rate code or the presence of negative meter readings).

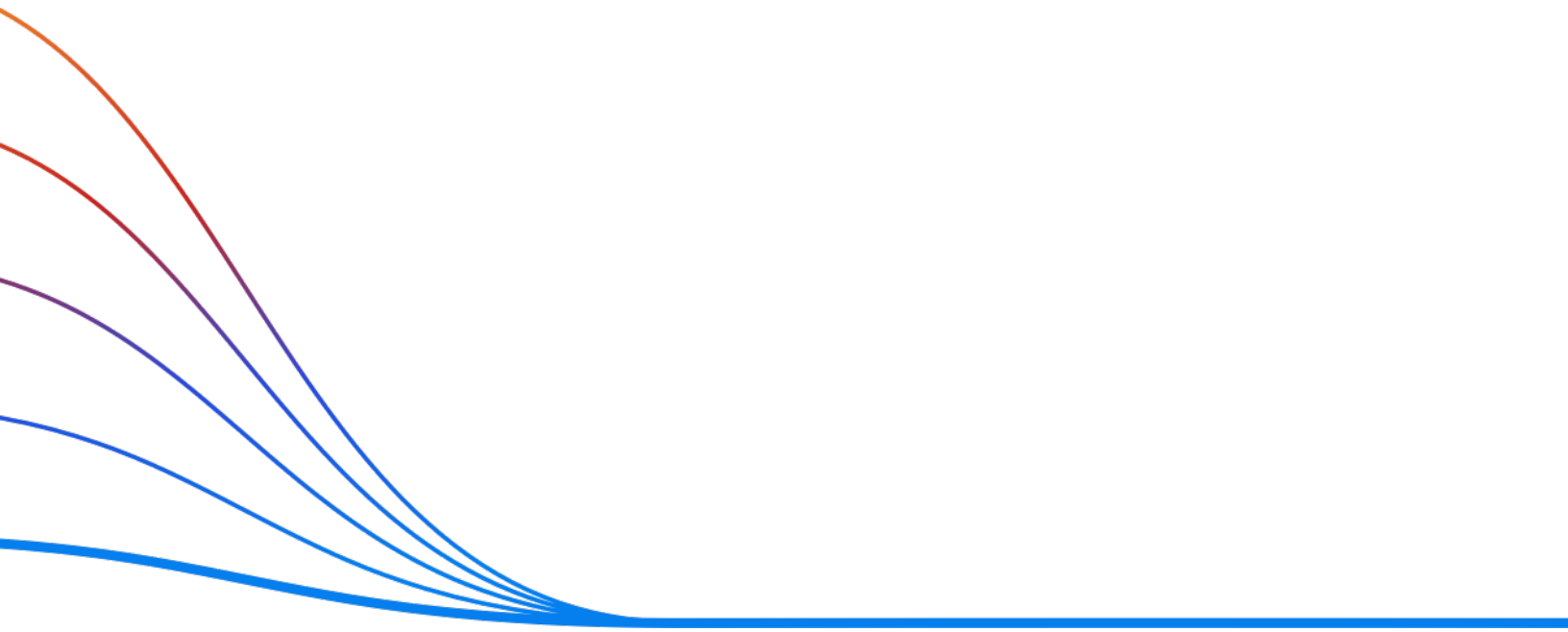
b. Had a total annual consumption in the baseline period greater than 50 MWh.
c. Had fewer than 329 days with at least one meter reading in the baseline period.
d. Had more than 15% of hours with null meter readings across the days in the baseline period with at least one meter reading.
e. Had fewer than 90% of days in the reporting period with at least one meter reading.
f. Had more than 15% of hours with null meter readings across the days in the reporting period with at least one meter reading.

2. Usage-based stratification parameters were computed for all meters.
3. The remaining meters were randomly split into two subgroups of equivalent size (50,000 meters each).
4. When testing against a territory-wide random sample, the first of these groups was always used as the random sample and the second group always served to furnish the city- or usage-based samples.
5. When testing samples with the same selection criteria, all qualifying meters from the first group were taken as the first sample residuals, and 3,000 random meters meeting the selection criteria were pulled from the second group.

The data and figures reported in Appendix B provide detailed results and figures for every test case summarized here.

RECURVE
SHAPE THE FUTURE OF ENERGY

Chapter 6: Quantifying Residuals and Variance in the Commercial Sector



I. Summary

This chapter details residual measurements from the difference of differences calculations specific to different business types in MCE service territory. These results provide a foothold for measuring the difference in consumption that could be expected due to COVID impacts and other exogenous factors when using the following measurement strategies:

1. No comparison group
2. A randomly selected comparison group of commercial customers
3. A comparison group of similar business type

Summary results from this experiment are provided in Table 7:

Table 7

	% Residual Total Savings			Hourly MAPE (%)	
NAICS Group	No Comparison Group	Random Comparison Group	Comparison Group of Peers	Random Comparison Group	Comparison Group of Peers
Administrative/Civil	-10.41	-6.09	-2.39	5.78	4.82
Automotive	-7.52	-6.67	2.24	7.08	3.88
Banks	-7.03	-8.12	0.31	8.04	3.61
Beauty	-59.60	45.11	1.63	32.69	2.78
Churches/Religious	-30.75	14.34	-2.22	12.49	4.11
Construction/Contractors	-10.06	-4.72	1.05	5.66	3.71
Fitness	-51.12	34.48	-2.68	31.54	6.33
Grocery/Convenience	-7.45	-7.12	1.49	7.31	4.05
Hotels/Lodging	-24.19	11.67	5.56	14.68	8.51
Medical_Offices	-17.30	3.89	3.79	8.17	6.35
Offices	-19.49	5.72	3.07	5.78	3.85
Real_Estate	-15.15	-0.13	0.04	2.78	2.25
Restaurants/Bars	-20.82	6.86	2.69	7.42	3.13
Retail	-21.41	5.05	-2.12	4.77	3.02
Schools	-42.56	29.58	4.64	22.24	11.59
Unassigned	-10.93	-4.09	0.57	3.83	1.46
Warehousing/Postal	16.03	-44.87	-27.09	43.88	33.13

Without a comparison group, the residuals of the total savings calculations ranged from -60% to 16% across different business types. Taking Offices as an example, without a comparison group we observe a 19.5% difference between forecasted and actual energy consumption using a standard OpenEEmeter savings calculation. Except for the Warehousing/Postal and Banks NAICS groups, incorporating a randomly selected comparison group consistently reduced the residual, often significantly. However, most sectors still exhibited a greater than 5% difference and a number of subsectors had residuals between 10 - 45%. When we introduced a comparison group consisting of business type peers, major improvements could be seen in every sector.

Looking at the Restaurant/Bars subsector, we see that without a comparison group, one would expect a residual of 21% in a savings measurement. A randomly selected comparison group reduces the residual to 7% and a reduction to under 3% is achieved by applying a comparison group of peers.

Similar improvements were observed in the hourly MAPE. When shifting from a randomly selected comparison group to a comparison group of peers, MAPE improved for all 17 NAICS groups. Continuing with the Restaurants/Bars example, MAPE is reduced from 7.4% to 3.1% in the hourly measurement.

One may expect that, as is done in stratified sampling, selection of comparison group meters based on common consumption characteristics would yield improvement over random sampling. For this to be the case particular baseline-period usage patterns would need to be identified that are strongly correlated with the energy consumption changes due to COVID. We have tested this possibility across a range of potential stratification parameters by first measuring meter-level COVID impacts and then gauging the degree to which many distinct usage parameters are correlated with changes in usage attributable to COVID. Results are given in Table 8.

Table 8

Parameter	Correlation with % COVID Impacts	Parameter	Correlation with % COVID Impacts
annual_kwh	-0.044	pct_winter	0.118
summer_kwh	-0.051	pct_winter_morn	0.006
summer_peak_kwh	-0.059	shldr_midday_restofday_ratio	-0.071
winter_kwh	-0.026	pct_baseload	0.078
winter_morn_kwh	-0.023	pct_variable	-0.078
shoulder_kwh	-0.049	pct_discretionary	-0.090
shoulder_midday_kwh	-0.068	pct_cooling	-0.057
cooling_kwh	-0.047	pct_heating	0.081
heating_kwh	0.018	evening_ramp_ratio	0.045
baseload_kwh	-0.002	summer_shldr_ratio	-0.018
variable_kwh	-0.070	utility_cost	-0.045
discretionary_kwh	-0.072	utility_cost_per_mwh	-0.034
evening_ramp	0.036	marginal_ghg	-0.043
pct_summer_peak	-0.082	marginal_ghg_per_mwh	0.006

While there are some interesting patterns in these results, the key takeaway is that none of these various 28 parameters, all computed from pre-COVID data, show a strong enough correlation with COVID impacts to warrant additional investigation.

II. Experimental Details

Similar to the residential experiments of Chapter 5, the following stepwise analysis was conducted to gauge the degree of residual in a % difference of differences calculation due to COVID.

1. Hourly OpenEEmeter 2.0 calculations were performed on all commercial meters in MCE territory using the timeline of Figure 1.

Meters were eliminated for which any of the following criteria were true:

   a. Were solar customers (identified by rate code or the presence of negative meter readings).
   b. Had a total annual consumption in the baseline period greater than 500 MWh.
   c. Had fewer than 329 days with at least one meter reading in the baseline period.
   d. Had more than 15% of hours with null meter readings across the days in the baseline period with at least one meter reading.
   e. Had fewer than 90% of days in the reporting period with at least one meter reading.
   f. Had more than 15% of hours with null meter readings across the days in the reporting period with at least one meter reading.

2. Remaining meters were randomly split into two equal subgroups of equivalent size (12,203 meters each).
3. The first of these groups was always used as the random sample.
4. Each NAICS group was also randomly split into two equivalent samples.
5. When computing the difference of differences calculation for a random sample vs. NAICS group, the first random sample was tested against the first NAICS subgroup. There will be a small degree of overlap between these two groups but this overlap should be 10% or less in every group except “Unassigned.”
6. When computing the difference of differences calculation for a NAICS group vs a peer group, the random samples from step 4 are tested against one another. There is no overlap between these groups.

Table 9 gives results (% Difference for the NAICS groups, % Difference of Differences for NAICS Group vs. Random and NAICS vs. Peers) for both the pre-COVID period and COVID periods.

Table 9

NAICS Group	NAICS Meter Count	Test:	Pre COVID NAICS % Diff	Pre COVID % Diff Diff	Pre COVID Hourly MAPE (%)	COVID NAICS % Diff	COVID % Diff Diff	COVID Hourly MAPE (%)
Administrative/Civil	2379	vs. Random	-0.11	0.05	2.31	-9.21	-6.09	5.78
		vs. Peers	-0.07	0.04	2.78	-11.60	-2.39	4.82
Automotive	878	vs. Random	-0.05	0.00	2.65	-8.63	-6.67	7.08
		vs. Peers	0.01	0.06	2.73	-6.40	2.24	3.88
Banks	188	vs. Random	-0.12	0.06	2.34	-7.18	-8.12	8.04
		vs. Peers	-0.09	0.03	2.61	-6.88	0.31	3.61
Beauty	952	vs. Random	-0.07	0.01	3.52	-60.42	45.11	32.69
		vs. Peers	-0.06	0.00	2.90	-58.78	1.63	2.78
Churches/Religious	565	vs. Random	-0.26	0.21	3.73	-29.64	14.34	12.49
		vs. Peers	-0.08	0.18	3.76	-31.86	-2.22	4.11
Construction/Contractors	939	vs. Random	-0.04	-0.01	2.68	-10.58	-4.72	5.66
		vs. Peers	-0.03	0.01	2.67	-9.54	1.05	3.71
Fitness	279	vs. Random	0.14	-0.20	4.36	-49.78	34.48	31.54
		vs. Peers	-0.92	-1.06	5.12	-52.46	-2.68	6.33
Grocery/Convenience	202	vs. Random	0.15	-0.21	3.28	-8.19	-7.12	7.31
		vs. Peers	0.00	-0.15	3.15	-6.70	1.49	4.05
Hotels/Lodging	274	vs. Random	-0.22	0.17	7.25	-26.97	11.67	14.68
		vs. Peers	-0.12	0.10	6.20	-21.41	5.56	8.51
Medical_Offices	1050	vs. Random	-0.01	-0.04	2.83	-19.19	3.89	8.17
		vs. Peers	-0.01	0.00	2.31	-15.40	3.79	6.35
Offices	1108	vs. Random	-0.03	-0.02	2.56	-21.02	5.72	5.78
		vs. Peers	-0.07	-0.03	2.30	-17.95	3.07	3.85
Real_Estate	2416	vs. Random	0.03	-0.08	1.58	-15.17	-0.13	2.78
		vs. Peers	0.05	0.03	1.39	-15.12	0.04	2.25
Restaurants/Bars	872	vs. Random	-0.01	-0.04	2.30	-22.16	6.86	7.42
		vs. Peers	0.08	0.09	1.24	-19.47	2.69	3.13
Retail	1325	vs. Random	-0.08	0.02	1.80	-20.35	5.05	4.77
		vs. Peers	-0.01	0.07	1.85	-22.47	-2.12	3.02
Schools	105	vs. Random	0.11	-0.16	9.99	-44.88	29.58	22.24
		vs. Peers	0.00	-0.11	10.39	-40.24	4.64	11.59
Unassigned	10712	vs. Random	-0.09	0.04	0.74	-11.21	-4.09	3.83
		vs. Peers	-0.03	0.06	0.85	-10.64	0.57	1.46
Warehousing/Postal	121	vs. Random	0.11	-0.16	5.41	29.57	-44.87	43.88
		vs. Peers	0.12	0.02	6.68	2.48	-27.09	33.13

Appendix C provides detailed results and figures for every test case summarized.

Methods Appendix

Savings Uncertainty Derivation

Given our savings equation:

$$S = \frac{o_{CG}}{m_{CG}} m_T - o_T \equiv F_{CG}\, m_T - o_T \tag{S1}$$

In S1, the comparison group terms are combined using the definition:

$$F_{CG} \equiv \frac{o_{CG}}{m_{CG}} \tag{S2}$$

We assume that observed meters have no uncertainty, $\sigma_{o_T} = 0$. Therefore, the uncertainty of S becomes the uncertainty of $F_{CG}\, m_T \equiv m_{T,c}$ and we can simply neglect o_T from the uncertainty calculations

$$m_{T,c} = \frac{o_{CG}}{m_{CG}} m_T \equiv F_{CG}\, m_T \tag{S3}$$

We start with the general equation for propagation of uncertainty with two variables

$$\sigma^2 = \left(\sigma_X \frac{\partial f}{\partial X}\right)^2 + \left(\sigma_Y \frac{\partial f}{\partial Y}\right)^2 + 2\, cov(X, Y) \frac{\partial f}{\partial X} \frac{\partial f}{\partial Y} \tag{S4}$$

The partial derivatives of Eq. S3 can be derived as:

$$\frac{\partial f}{\partial X} \equiv \frac{\partial m_{T,c}}{\partial F_{CG}} = m_T \tag{S5}$$

$$\frac{\partial f}{\partial Y} \equiv \frac{\partial m_{T,c}}{\partial m_T} = F_{CG} \tag{S6}$$

By substituting Eqs. S5-6 into Eq. S4:

$$\sigma^2_{m_{T,c}} = \left(\sigma_{F_{CG}} m_T\right)^2 + \left(\sigma_{m_T} F_{CG}\right)^2 + 2\, cov(F_{CG}, m_T)\, F_{CG}\, m_T \tag{S7}$$

If we then divide Eq. S7 by Eq. S3 squared:

$$\frac{\sigma^2_{m_{T,c}}}{m^2_{T,c}} = \frac{\left(\sigma_{F_{CG}} m_T\right)^2}{m^2_{T,c}} + \frac{\left(\sigma_{m_T} F_{CG}\right)^2}{m^2_{T,c}} + 2\, cov(F_{CG}, m_T)\, \frac{F_{CG}\, m_T}{m^2_{T,c}} \tag{S8}$$

$$\left(\frac{\sigma_{m_{T,c}}}{m_{T,c}}\right)^2 = \left(\sigma_{F_{CG}} \frac{m_T}{F_{CG}\, m_T}\right)^2 + \left(\sigma_{m_T} \frac{F_{CG}}{F_{CG}\, m_T}\right)^2 + 2\, cov(F_{CG}, m_T) \frac{F_{CG}\, m_T}{(F_{CG}\, m_T)^2} \tag{S9}$$

Some algebraic simplification:

$$\sigma_{m_{T,c}} = |m_{T,c}| \sqrt{\left(\frac{\sigma_{F_{CG}}}{F_{CG}}\right)^2 + \left(\frac{\sigma_{m_T}}{m_T}\right)^2 + 2\frac{cov(F_{CG}, m_T)}{F_{CG}\, m_T}} \tag{S10}$$

Substituting σ for ϵ to match the nomenclature of the manuscript, understanding that σ is the standard deviation and ϵ represents the uncertainty which is a function of standard deviation, and the previous conclusion that $\sigma_S \equiv \sigma_{m_{T,c}}$:

$$\epsilon_S = |m_{T,c}| \sqrt{\left(\frac{\epsilon_{F_{CG}}}{F_{CG}}\right)^2 + \left(\frac{\epsilon_{m_T}}{m_T}\right)^2 + 2\frac{cov(F_{CG}, m_T)}{F_{CG}\, m_T}} \tag{S11}$$

The express purpose of using %-diff-of-diff is to correct the model based upon a comparison group. Eq. S11 does not take this into consideration. Because of this $m_{T,c}$ will be substituted in for m_T with the expectation that the comparison group will reduce the error in the model and this will be captured in the baseline using the RMSE of the corrected model.

$$\epsilon_S = |m_{T,c}| \sqrt{\left(\frac{\epsilon_{F_{CG}}}{F_{CG}}\right)^2 + \left(\frac{\epsilon_{m_{T,c}}}{m_{T,c}}\right)^2 + 2\frac{cov(F_{CG}, m_T)}{F_{CG} m_{T,c}}} \tag{S12}$$

$$\epsilon_S = \sqrt{\left(\epsilon_{F_{CG}} \frac{m_{T,c}}{F_{CG}}\right)^2 + \epsilon_{m_{T,c}}^2 + 2\, cov\left(F_{CG}, m_{T,c}\right)\frac{m_{T,c}}{F_{CG}}} \tag{S13}$$

One further simplification can be performed by substituting Eq. S3 into Eq. S13

$$\epsilon_S = \sqrt{\left(m_T\, \epsilon_{F_{CG}}\right)^2 + \epsilon_{m_{T,c}}^2 + 2\, cov\left(F_{CG}, m_{T,c}\right) m_T} \tag{S14}$$

This is Eq. 18 from the manuscript. $\epsilon_{F_{CG}}$ can be further broken-down. In practice F_{CG} is calculated as the weighted average of o_{CG}/m_{CG}. This means that the uncertainty of the F_{CG} must take into account all of the model uncertainties as well as the variance of all these values when they are aggregated.

We will need to take a small detour to derive the formula for $\epsilon_{F_{CG}}$. We'll start with the definition of variance.

$$Var(X) = E(X^2) - E(X)^2 \tag{S15}$$

Here X is taken to be the population average of a set of individual "true" values X_i, each of which has been measured with some uncertainty as $x_i \pm \epsilon_i$, and E represents taking an expectation value. (In our particular case, the measurements are the F_{CG_i}, and we are trying to find the average value F_{CG}.) Then $E(X)$ is just the average of the x_i values, while $E(X^2)$ is the average of the expected values X_i^2, which is given by Eq. S16.

$$E\left(X_i^2\right) = Var(X_i) + E(X_i)^2 = \epsilon_i^2 + x_i^2 \tag{S16}$$

$$Var(X) = E\left(\epsilon^2 + x_i^2\right) - E\left(x_i^2\right) = E(\epsilon_i^2 + (x_i - \mu)^2) \tag{S17}$$

Here, μ is the average of the x_i and we have used the fact that $E((x_i - \mu)^2) = E(x_i^2) - \mu^2$, which is straightforward to prove.

Converting Eq. S17 back to our original notation and recalling that we are calculating a weighted average, the F_{CG} uncertainty is shown to be the square root of the variance, multiplying by the t-statistic, and accounting for the proper degrees of freedom to create a confidence interval at a chosen confidence level (which we will typically take to be 95% in this paper).

$$\epsilon_{F_{CG}} = \sqrt{\sum \left(w_i \left(\epsilon_{F_{CG_i}}^2 + \left(F_{CG_i} - F_{CG} \right)^2 \right) \right)} \frac{t_{\frac{\alpha}{2}, N-1}}{\sqrt{N}} \tag{S18}$$

Here, w_i is the weight of each meter in the comparison group normalized such that the sum is 1, N are the total number of meters in the comparison group, t is the aforementioned t-statistic, and α is the level of significance. For IMM, the weights are the meter's model value for a given hour divided by the sum of all meter's model values in the comparison group. If there are negative weights, they are set to zero and then the weights are renormalized to 1. For all clustering methods, the weights are 1/N making this an unweighted average. For the uncertainties of the individual $\epsilon_{F_{CG_i}}$, we return to Eq. S4 which requires the partial derivatives of Eq. S2. We have added the subscript i as necessary because Eq. S2 is being applied to individual meters at this point.

$$\frac{\partial f}{\partial X} \equiv \frac{\partial F_{CG_i}}{\partial o_{CG_i}} = \frac{1}{m_{CG_i}} \tag{S19}$$

$$\frac{\partial f}{\partial Y} \equiv \frac{\partial F_{CG_i}}{\partial m_{CG_i}} = -\frac{o_{CG_i}}{m_{CG_i}^2} \equiv -\frac{F_{CG_i}}{m_{CG_i}} \tag{S20}$$

Since we've already made the assumption that the observed meters have no uncertainty this means that the uncertainty of F_{CG_i} can be shown by Eq. S21.

$$\epsilon_{F_{CG_i}}^2 = \left(-\frac{F_{CG_i}}{m_{CG_i}} \sigma_{m_{CG_i}} \right)^2 \equiv \left(F_{CG_i} \frac{\sigma_{m_{CG_i}}}{m_{CG_i}} \right)^2 \tag{S21}$$

We have further simplified the equation by removing the negative sign because we are taking the square, and we can use the definition of F_{CG}, Eq. S2, to simplify the equation further. We then substitute back into Eq. S18 resulting in Eq. S22 while also substituting σ for ϵ.

$$\epsilon_{F_{CG}} = \sqrt{\sum \left(w_i \left(F_{CG_i} \frac{\epsilon_{m_{CG_i}}}{m_{CG_i}} \right)^2 + w_i \left(F_{CG_i} - F_{CG} \right)^2 \right)} \frac{t_{\frac{\alpha}{2}, N-1}}{\sqrt{N}} \tag{S22}$$

The only piece left underived is the uncertainty of the model which is the model's RMSE in the baseline period in where it was fit. The resulting uncertainty, $\epsilon_{F_{CG}}$, from Eq. S22 is then solved and can be substituted back into Eq. S14 for the final savings uncertainty.