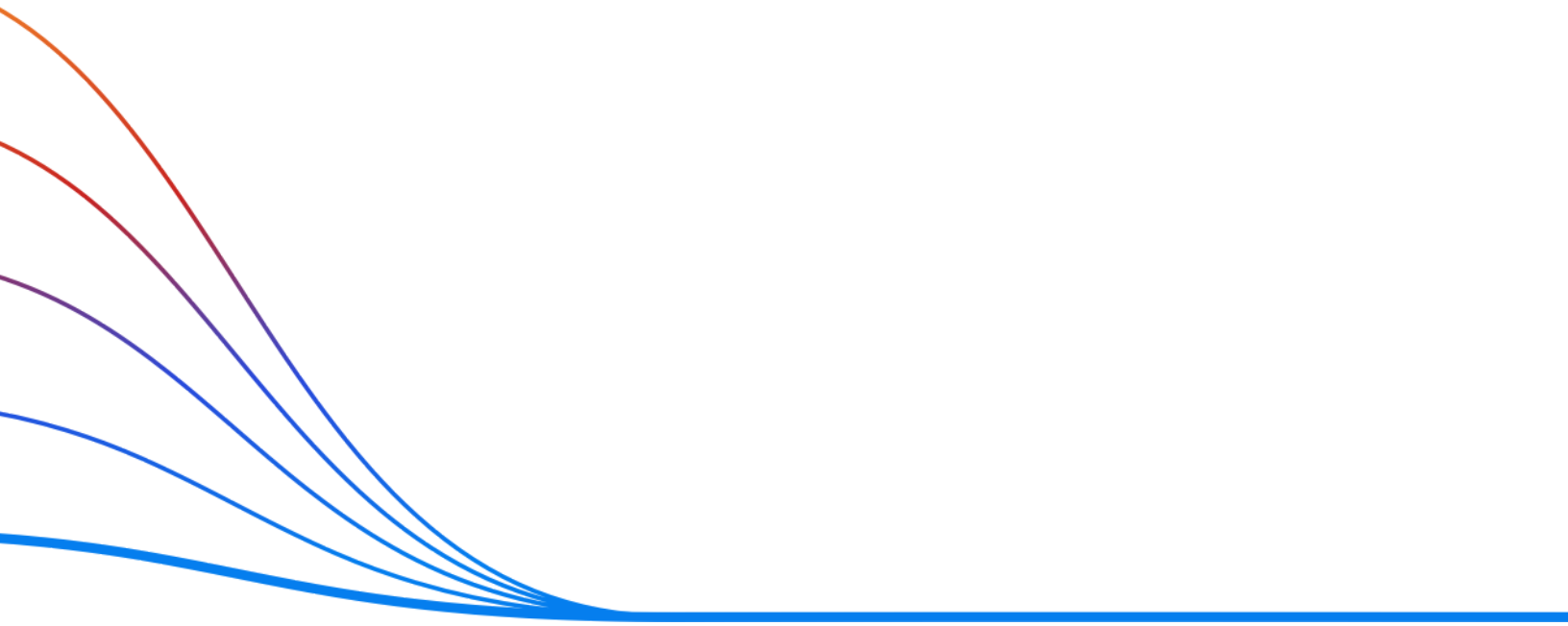
RECURVE

SHAPE THE FUTURE OF ENERGY

Comparison Groups for the COVID Era and Beyond



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.

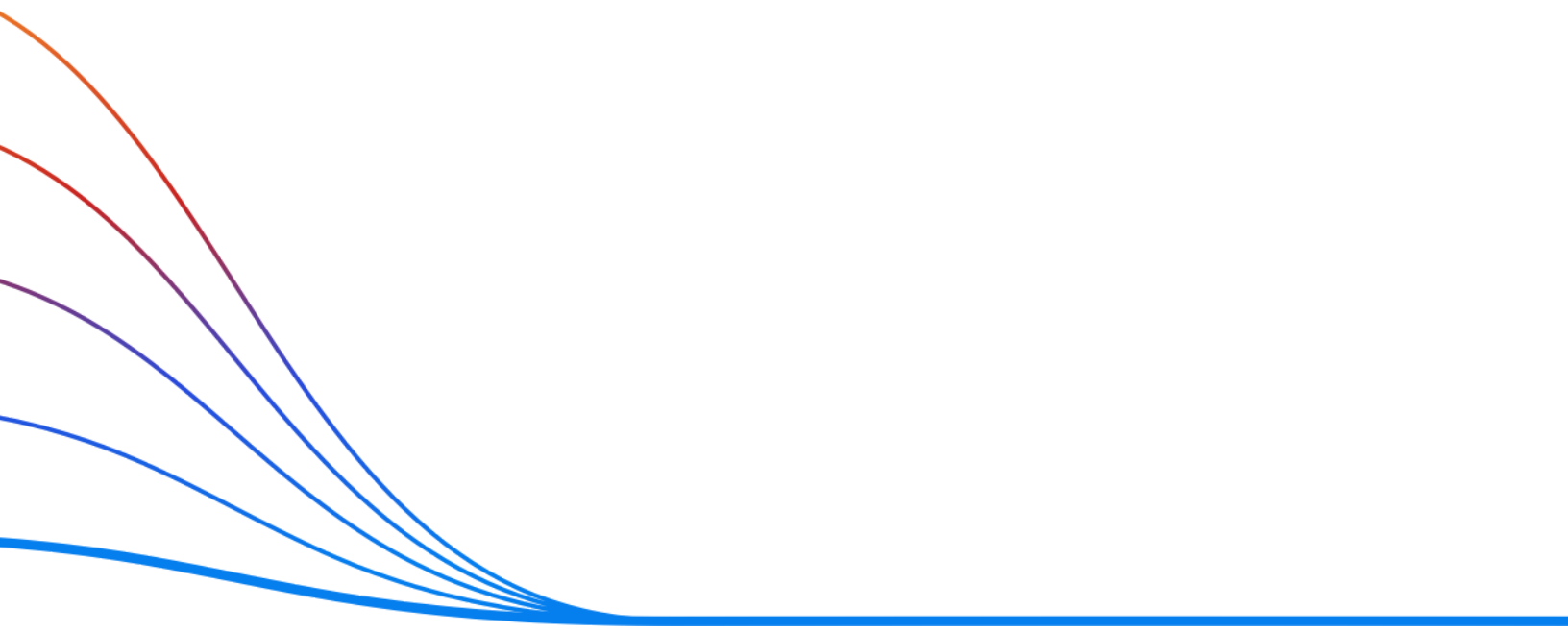
Recurve would also like to thank MCE for supporting this project by providing secure access to data. Without secure data-sharing partnerships between utilities/energy providers and the demand side service industry, the next generation of programs capable of fighting climate change, enhancing grid resilience, 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



Executive Summary

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 adjust the gross 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 that 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 that comparison groups are needed to adjust for significant exogenous effects on building energy consumption.

In the guidance that follows, many of the concepts of comparison group methodologies will seem familiar. However, the particular use case is distinct from the multifaceted objectives of final program evaluations. 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 particular 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 to

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

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 particular 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 CalTRACK[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 CalTRACK 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 buildings likely to respond similarly to exogenous factors as would be expected of buildings enrolled in a demand-side energy program. 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] http://docs.caltrack.org

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 recommendations provided 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.

The following recommendations flow from experience measuring the impacts of dozens of demand-side programs with both monthly and AMI data and from research results presented throughout the next five chapters and supporting appendices. The chapters that follow provide much more explanation, rationale, and data, and we encourage readers to explore this content.

Grid Methods Synopsis

1. Identify program-eligible participants

    1.1. All demand-side energy programs will have eligibility rules. Some are based on customer-specific designations, such as low-income or hard-to-reach customers. Others are based on sector, such as commercial, agricultural, residential, or industrial. Some programs may be intended for non-solar customers, while others may require that a customer have solar PV. Yet other programs might restrict participation based on energy consumption characteristics, such as high peak usage or annual usage within a certain range. These eligibility rules are valuable because they tend to organize customers into classes that respond similarly to exogenous events.

2. Limit comparison group to eligible customers that meet program requirements. Identifying eligibility is also the first step to defining a relevant comparison pool from which a comparison group will ultimately be formed.

    2.1. Fit a CalTRACK 2.0 model on all eligible program participants prior to program launch. This model will uncover incomplete or missing data, erratic energy consumption patterns, and potential for higher savings. If program optimization techniques are applied, such as selecting targeted customers based on energy consumption profiles, customers who fit these criteria can be proportionally sampled as described in Chapter 3 in order to more specifically anticipate the likely program enrollees.

    2.2. Remove outlier customers from the comparison group sampling pool.

    2.3. Array all eligible customers by annualized consumption by fitting the baseline CalTRACK model to weather conditions of the baseline year.

    2.4. Remove customers with daily baseline CVRMSE values in excess of 1.0 or more rigorous thresholds depending on the program's need.

    2.5. Remove any remaining customers failing to meet program eligibility criteria.

3. Determine Comparison Group Requirements

    3.1. Annual savings that rely on daily or monthly savings calculations require different comparison group selection criteria than marginal hourly savings.

    3.2. Small treatment groups, irrespective of the granularity of savings calculations, require different comparison group selection criteria than large treatment groups.

4. Random Selection of Comparison Group from Population

    4.1. For some programs a sufficient comparison group can be formed via random selection from within an eligible comparison pool. The random selection should be made after filtering for program and other eligibility requirements.

    4.2. A comparison group selected randomly from an eligible population must minimize the potential for sampling error by selecting a large pool of non-participants. The assumption with this type of comparison group is that the treated customers are experiencing exogenous factors in the same way as the larger population. In this case, the randomly selected comparison group is expected to be representative of the larger population and is thus a suitable basis for calculating exogenous effects. Comparison group sizing is described in greater detail in Chapter 2.

5. Random Selection of Comparison Group from Sub-Population

    5.1. If a treated group is not expected to reflect the usage patterns or bear the impact of exogenous factors the same as the population as a whole, the comparison group must

be designed to reflect the treated group rather than the population. In this case, the sampling error observed is between the comparison group and the treated group rather than the comparison group and the population. Chapters 5 and 6 and Appendices B and C give more detail on quantifying and minimizing COVID-related residuals in the Residential and Commercial sectors.

5.2. A treated group may be drawn from a targeted subset of the program eligible population. For example, a program may target customers with usage patterns substantially different from the program eligible population, such as those whose energy use peaks during peak evening hours. In this case, it will be important to draw a sample of non-participants from within the distribution of targeted participants.

5.3. Targeting parameters will skew the distribution of a treated group away from the general population. The comparison group should attempt to replicate this skewness (as well as the probable kurtosis resulting from optimization strategies).

5.4. If the treated group is likely to be large and normally distributed amongst the targeted population, a large and normally distributed comparison group drawn from within the same population will be the best way to achieve the desired similarity.

6. Selection of Comparison Group from within Stratified Sample of Sub-Population

6.1. If a treated group is substantially different from the comparison group selected, the sub-population may be resampled to select a comparison group more similar to the treatment group.

6.2. Resampling should only occur once enrollment in the program has reached a sufficient level to support stratified sampling from the broader population of eligible non-participants. Chapter 3 provides a detailed procedure for conducting and optimizing stratified sampling.

6.3. Multiple approaches to binning based on consumption parameters are acceptable. However, it should be noted that stratified sampling will not solve certain problems such as categorical bias, for example, where certain business sectors are differently affected by exogenous variables than other business sectors. Chapter 3 provides a decision framework to identify where random, proportional, or stratified sampling should be conducted.

7. Creation of Comparison Group Vintages and Difference of Differences

7.1. Once a comparison group has been created, the baseline period of the comparison group must be aligned temporally with the baseline period of the participating customers.

7.2. Where programs enroll customers over a period of time longer than 30 days, the comparison group must be rebaselined for each month of enrollment and a new vintage created that is assigned to a monthly cohort of enrolled participants.

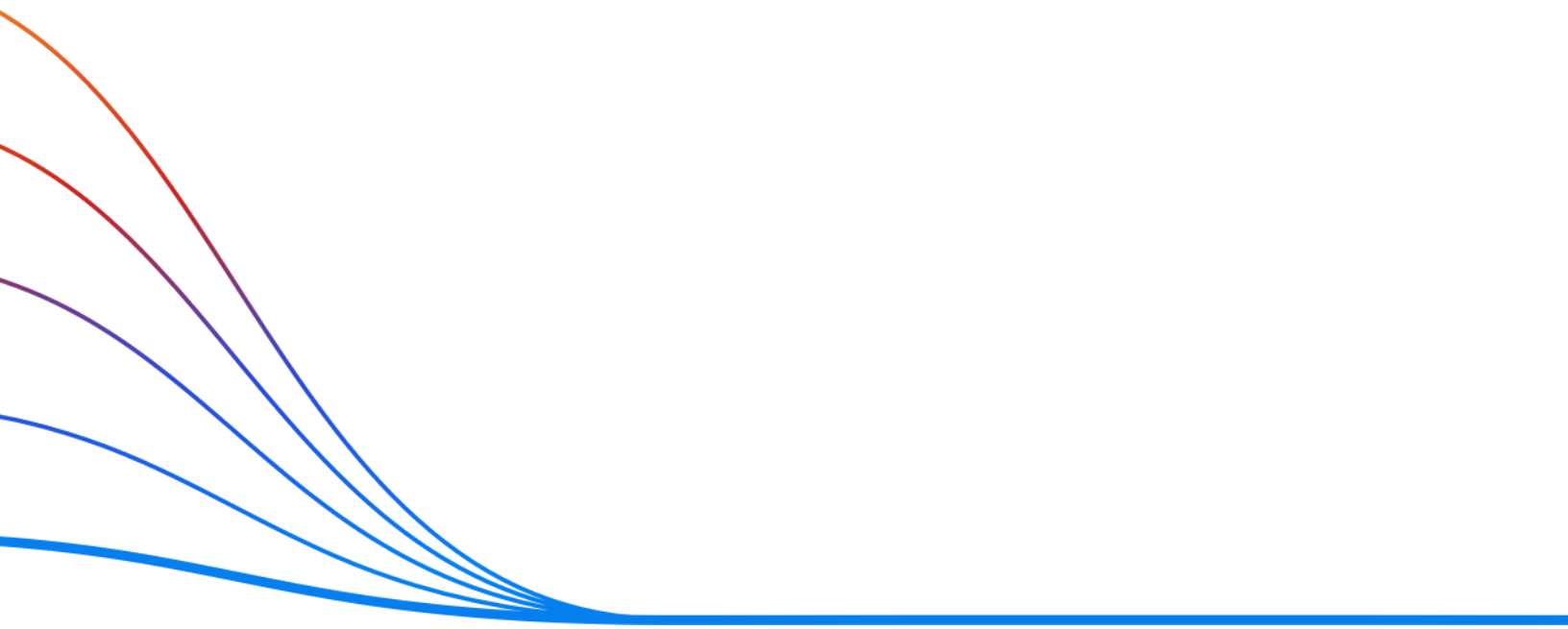
7.3. For each monthly cohort of participants, calculate a difference of differences of percentage savings between the treated customers and the associated vintage of the comparison group.

7.4. The difference of differences in percentage terms can be multiplied by the raw total for the purposes of aggregation of multiple treated cohorts.

7.5. The difference of differences calculation should be applied to the model counterfactual for the determination of savings. More detail is provided in Chapter 4 on conducting the difference of differences savings calculations.

RECURVE
SHAPE THE FUTURE OF ENERGY

Chapter 2: Key Analyses and Results That Inform Recommendations



I. 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 CalTRACK 2.0 hourly calculations for each meter using the following baseline and reporting period timeline:

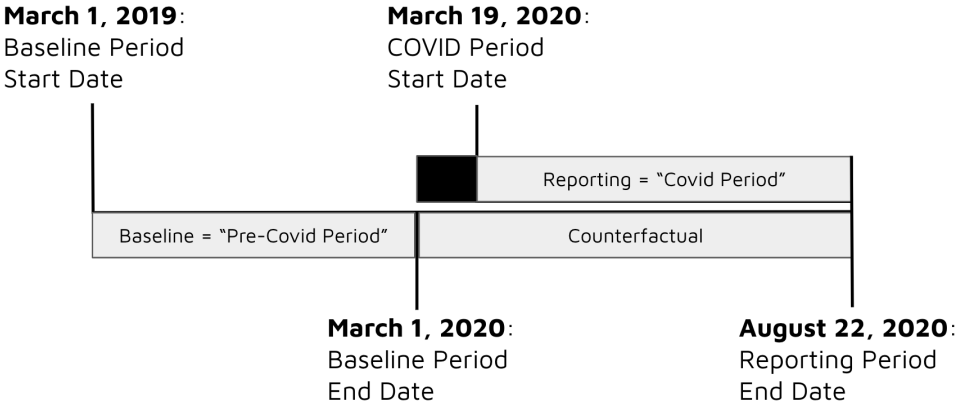


Figure 1: Metering timeline for the calculation of COVID impacts. This timeline is used for many of the comparison group tests described throughout this document.

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 CalTRACK 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.

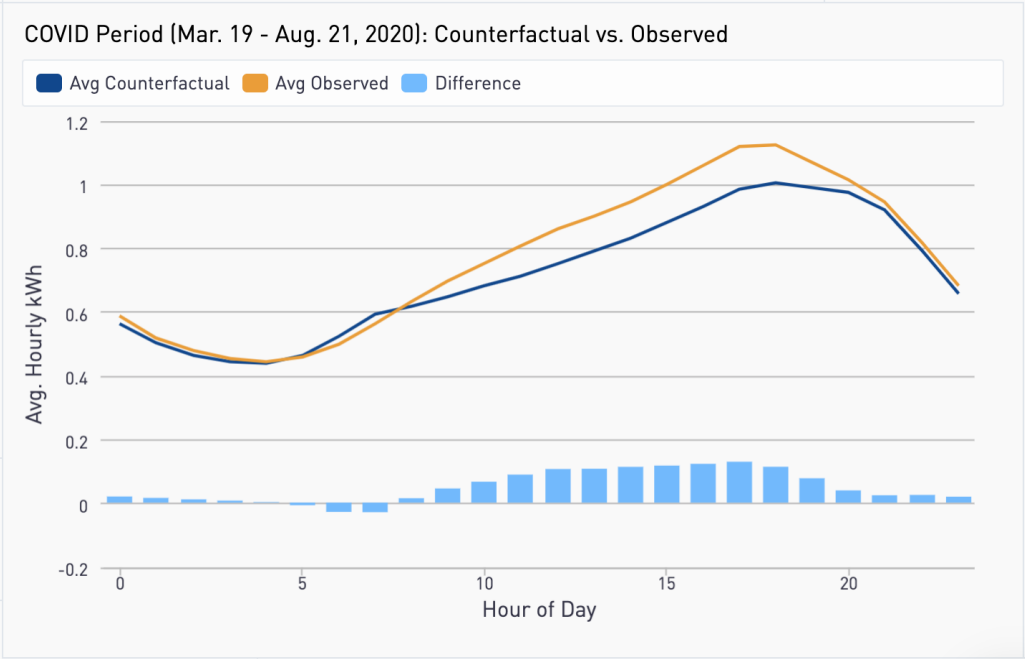


Figure 2: Average observed and counterfactual daily load shapes for MCE non-solar Residential customers during the COVID period.

We measure a total increase in consumption of 7.9% due to COVID, with the majority 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.

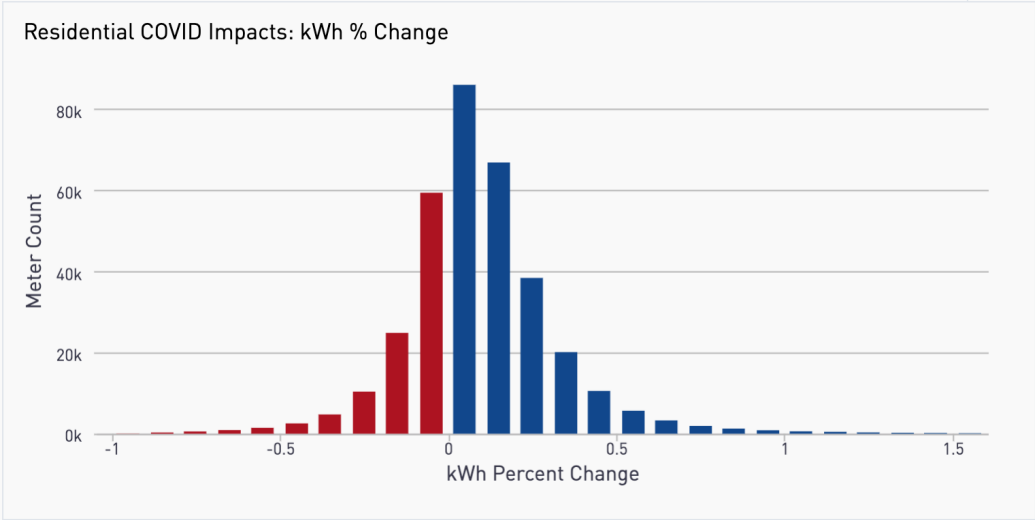


Figure 3. Distribution of the percent change in electricity consumption due to COVID for MCE non-solar Residential customers.

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.

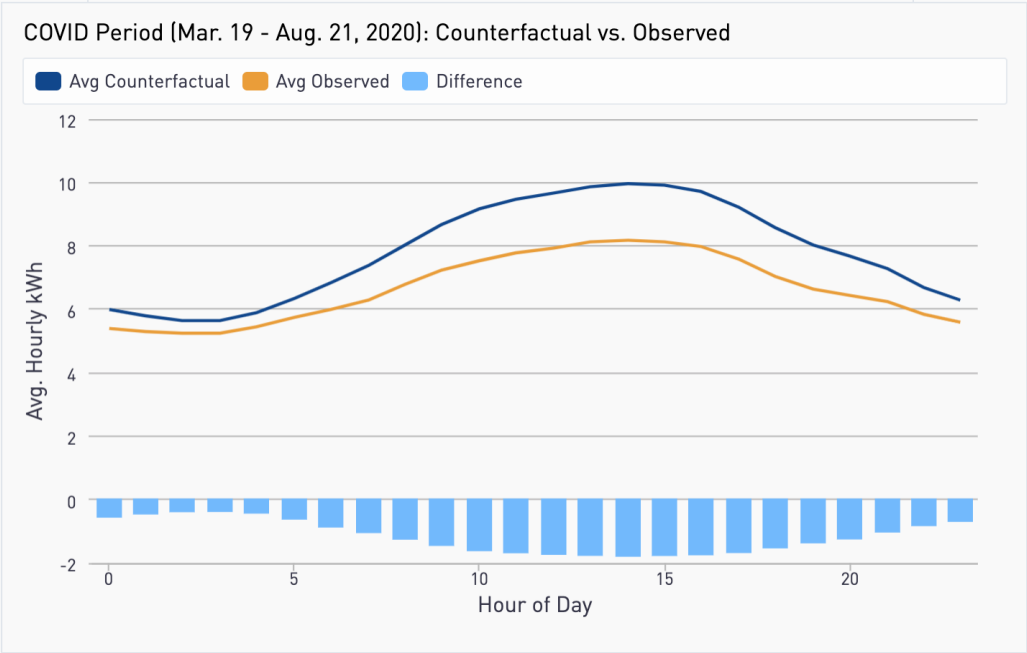


Figure 4: Average observed and counterfactual daily load shapes for MCE non-solar Commercial customers during the COVID period.

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]

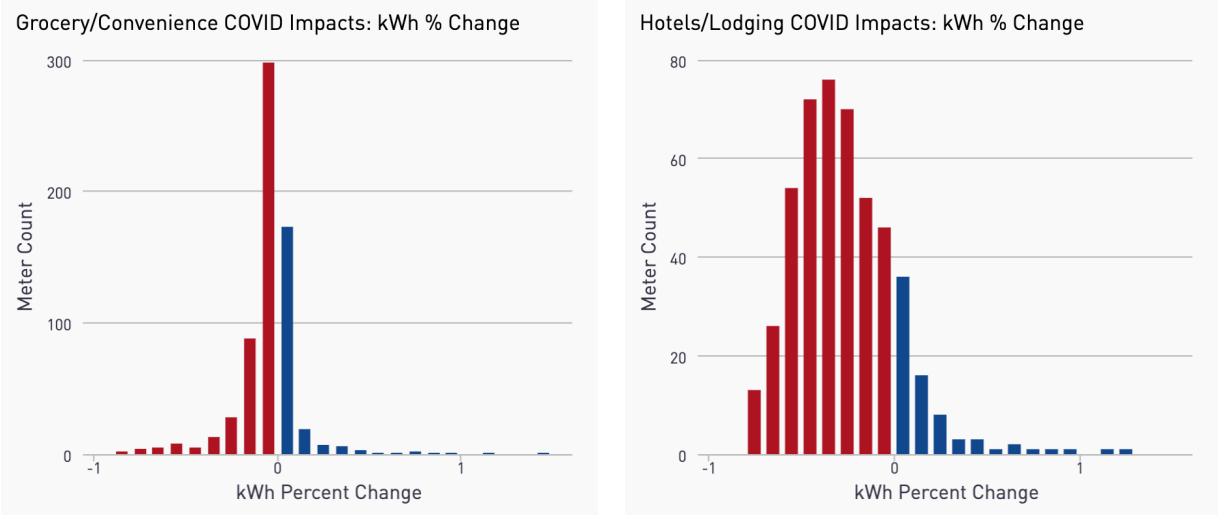


Figure 5: Distribution of the percent change in electricity consumption due to COVID for Grocery and Convenience stores (left) and Hotels and Lodging facilities (right).

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 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).

[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.

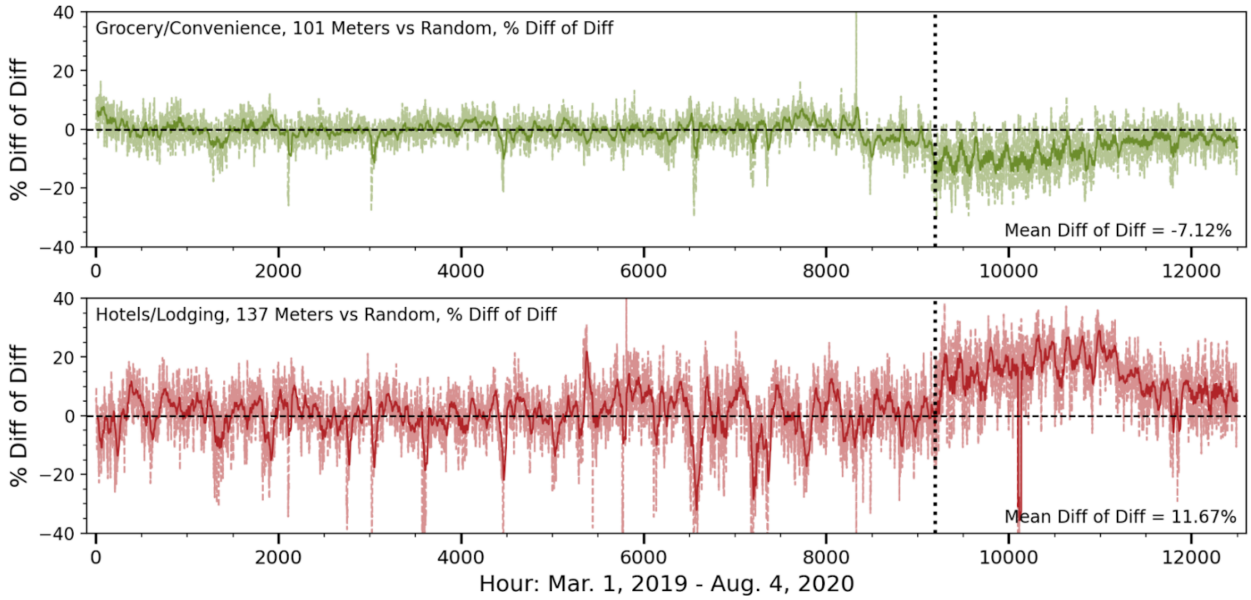


Figure 6: Difference of differences calculations throughout the baseline and COVID periods for the Grocery/Convenience segment vs. a random sample of commercial customers (top) and the Hotels/Lodging sector vs. a random sample of commercial meters (bottom). The dotted line indicates the start of the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

As detailed in Chapter 6/Appendix C, with data on business type, comparison groups can be formed that are capable of neutralizing 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 particular consumption characteristics that are adequately predictive of COVID impacts to eliminate the business type differences we observe.

II. 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 actually 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 CalTRACK 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 CalTRACK model.

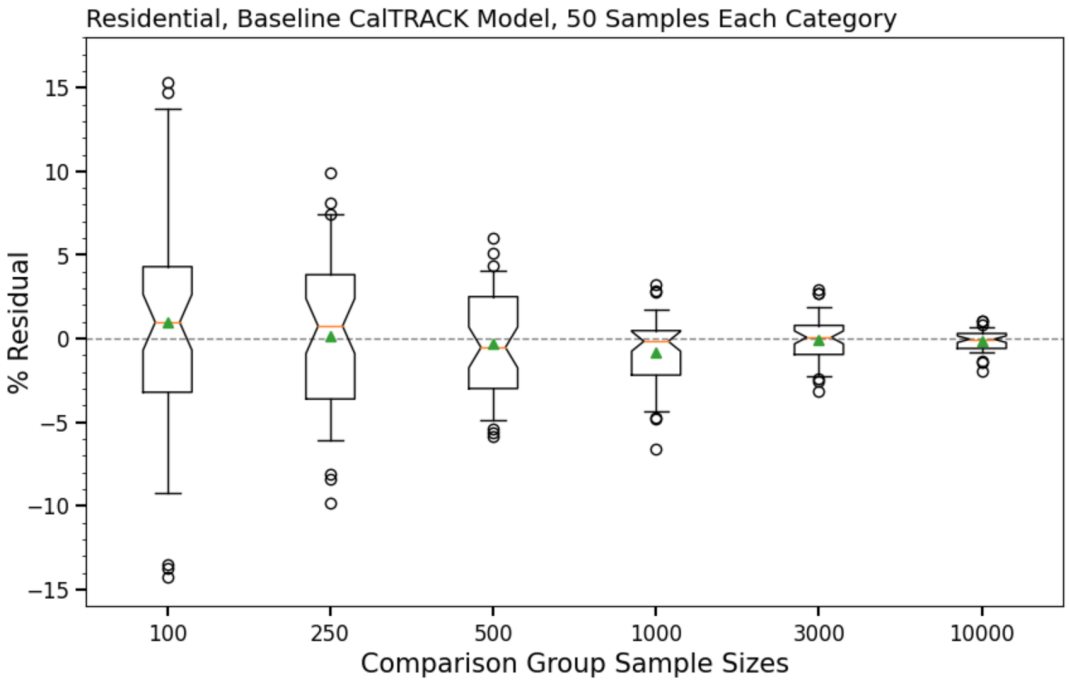


Figure 7: Box and whisker plot showing the distribution of residual error introduced by different comparison group sample sizes for Residential customers in MCE’s service territory. Each box represents the interquartile range and the whiskers represent the 0.05 - 0.95 probability range. Outliers are 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 are capable of reducing 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.

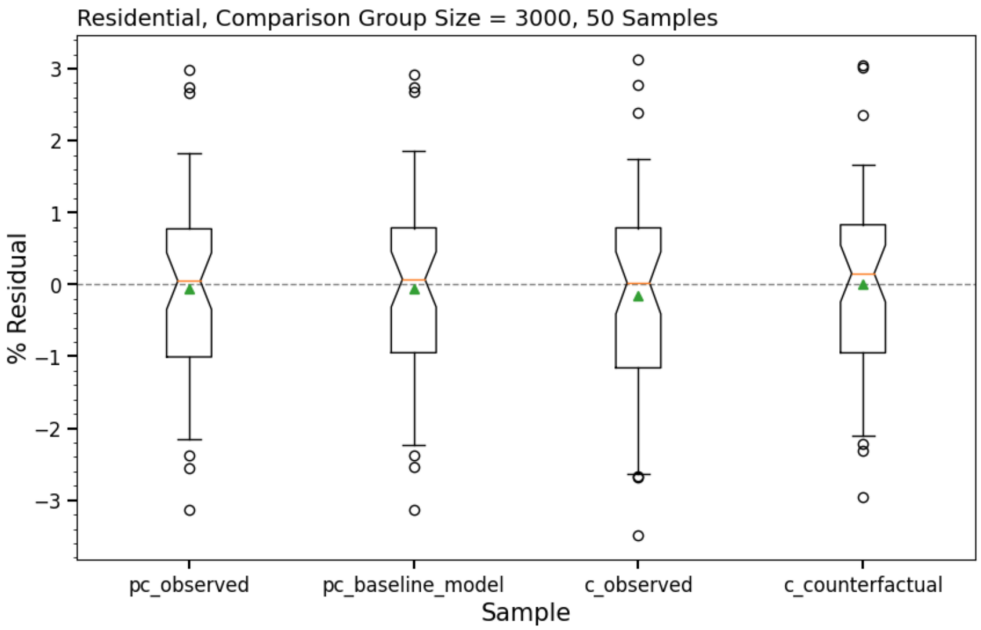


Figure 8: Box and whisker plot showing the variability present across each component of the difference of differences calculation for the 3000 sample size. “pc” indicates pre-COVID and “c” indicates COVID.

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.

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 CalTRACK

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

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.

III. 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.

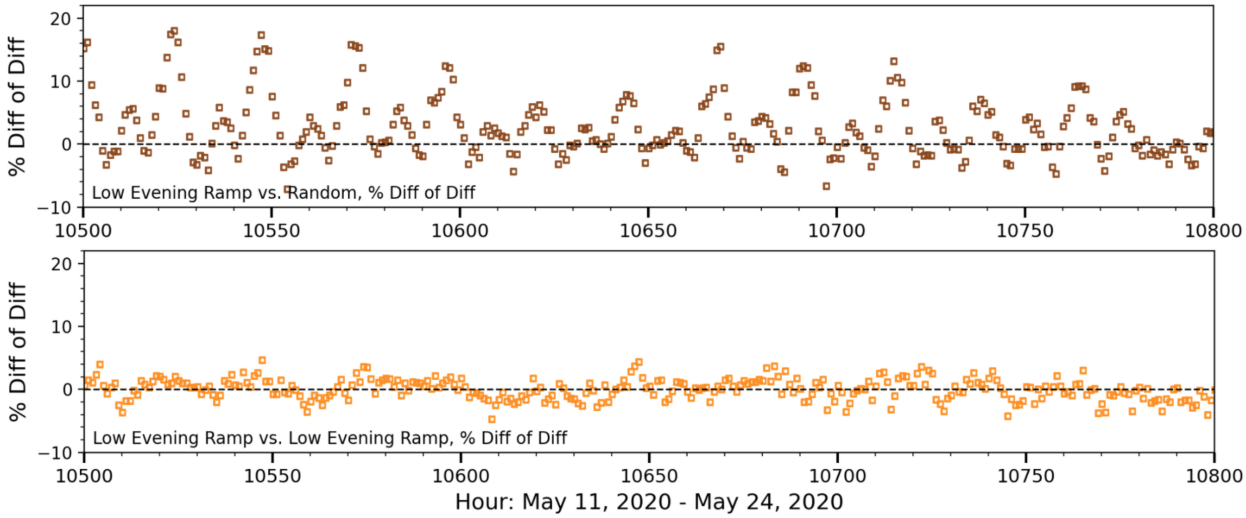


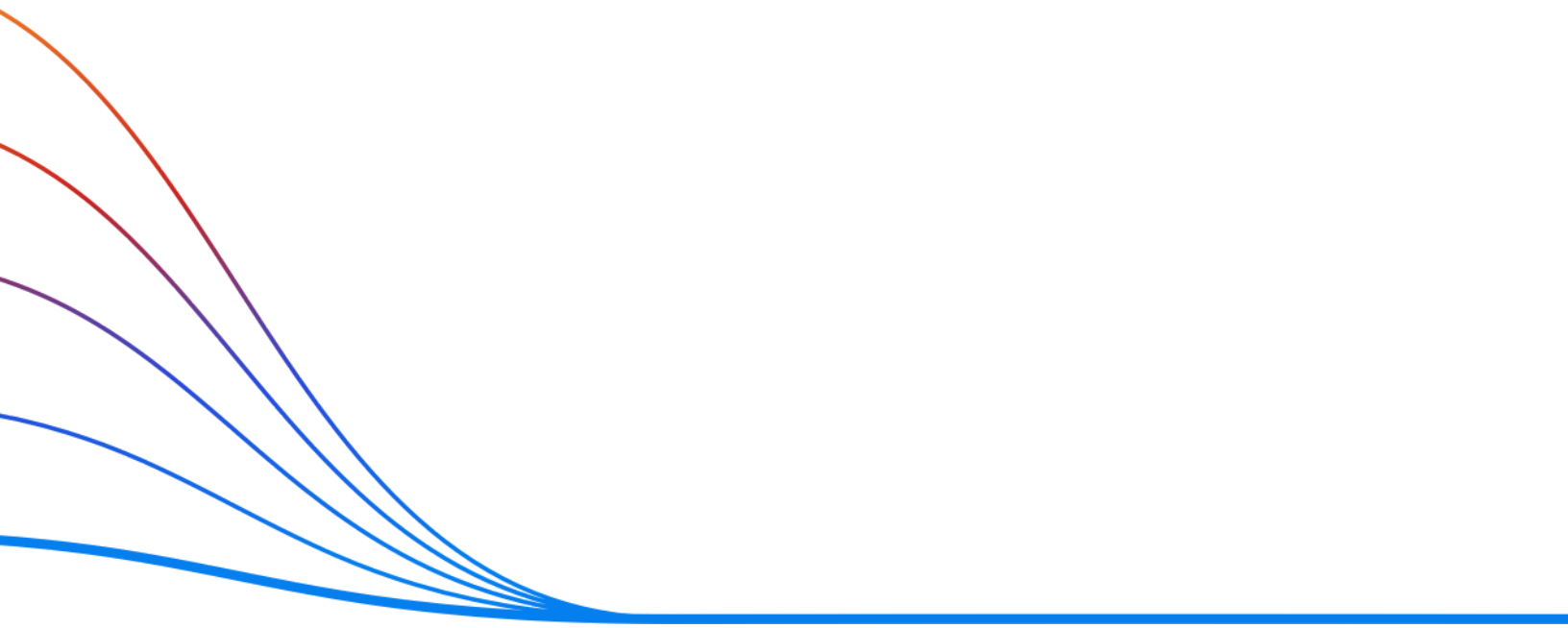
Figure 9: Residuals in a difference of differences calculation. Top - sample of residential meters with low evening ramp vs. a random sample of residential customers. Bottom - sample of residential meters with low evening ramp vs. a sample of residential customers that have similar load characteristics.

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: Recommended 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 sampling based either on a program's forecasted participation or on the population of actual 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 complex or computationally expensive to implement.

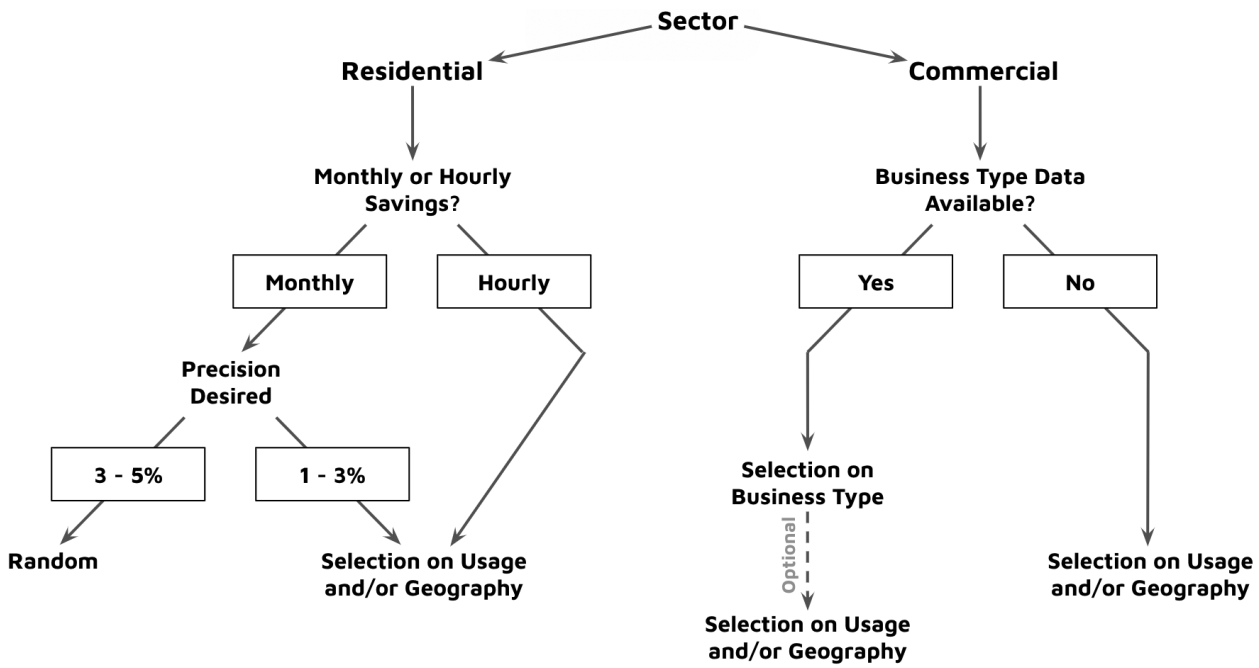
In addition to these conditions, the following recommendations are rooted in the testing of many different comparison group scenarios, which are detailed in Chapters 5 and 6 and complimentary appendices B and C.

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.

In particular, 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 is in contrast to 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, Chapter 5/Appendix B shows that hourly measurements benefit from a comparison group selected based on additional criteria, including geographic location and usage characteristics. 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. Program Forecasting Stage

It will be advantageous for many meter-based pay-for-performance programs to have a comparison group established at the outset based on a forecasted population of participants. The forecast-based comparison group will help get the program on track for savings calculations and aggregator performance payments. As it cannot be known in the forecasting phase which specific customers will ultimately participate,[5] some common comparison group strategies, including individual site-based matching and future participants, will not be possible at the outset.

At this point, according to the decision tree above, either a random sampling or proportional sampling scheme can be deployed. In either case, it will be important that the comparison pool is limited to customers that meet the same data sufficiency and program eligibility considerations required for participating customers. For instance, if the program requires participating customers to have a minimum annual usage of 1 MWh and no solar PV system, then the comparison pool should be screened for these criteria as well.

In the Residential sector, proportional sampling can be conducted on the basis of both geographic location and usage characteristics expected for program participants. If an efficiency program in Arizona wished to target high air conditioning users in the low elevation climate regions, proportional sampling could be done from these geographic areas and weighted to populations with high cooling loads observed at the meter. Chapter 5/Appendix B demonstrates how both geographic and usage-based sampling can reduce residuals in the calculation of load impacts, including on an hourly basis.

[5] Note that for opt-out program designs it can be known in advance which customers will be included. Some of these programs can be administered as a randomized control trial (RCT). Examples and literature of energy efficiency programs conducted as RCTs are well documented and are not the focus of this effort.

In the commercial sector, we observe that business type is the most reliable predictor of population-level consumption changes due to COVID. Therefore, our recommendation is to utilize business type (via NAICS code or other data sources) information as a selection parameter in the formulation of a comparison group wherever these data are available. For forecasting, the program administrator and aggregator should assess likely participating business types and formulate a comparison group based on the outcome of this process. Where business type information for the comparison pool is not available, sampling based on expected geographic location and customer usage characteristics can help to identify the most relevant comparison group. However, due to the high degree of variance in COVID impacts by economic sub-sector, where business type data are not available, program administrators and aggregators will face increased risk. In these cases, serving a wide variety of customers in the program can help mitigate risk.[6]

III. Program Implementation Stage

After a program enters its enrollment phase, an actual participant group will emerge. Depending on the circumstances of the in-field program, this group may or may not align well with the forecast. Differences could emerge related to business type, geography, usage characteristics, or other important factors. In these cases, program administrators and aggregators may wish to reformulate the comparison group to better represent the actual program participation. This step can mitigate risk for both parties associated with a mismatched comparison group.

With COVID impacts driving the largest non-program changes in energy usage and business type being the best predictor of these changes, comparison group resampling for Commercial programs should be done first to ensure that the proportion of business types in the comparison group matches that of the treatment group. Given sufficient meters in the comparison pool, the comparison group can be further refined to best capture geographic and/or usage characteristics per the stratification methods detailed in the next section.

For the Residential sector, we recommend that resampling be done based on the stratified or meter matching sampling approach described below.

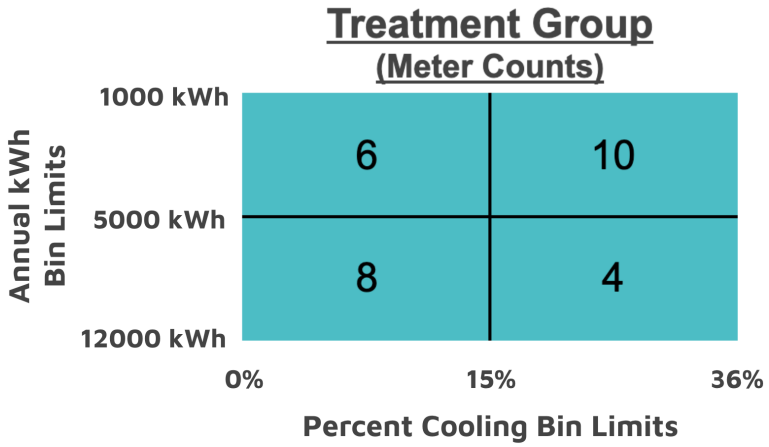
[6] Several strategies were explored in this research to select comparison groups from usage characteristics alone that could reliably eliminate COVID impacts in the Commercial sector. One such strategy that showed promising results was to use CalTRACK models to estimate consumption changes due to COVID at an individual-meter level for trial treatment groups and comparison pools. The percent change in consumption for each meter could then be used directly as a selection parameter. While this method appeared capable of reducing residuals, we ultimately did not pursue recommendations around this approach for a couple reasons. First, as time goes on the pre-COVID baseline period becomes further in the past, which will inherently reduce the size of the comparison pool and leave fewer program-eligible customers. Increased time between pre-COVID period and program enrollment will also make the baseline period less and less relevant. Second, the extra step of calculating COVID-impacts for each meter adds a great deal of complexity and can delay the process of finalizing a comparison group. For these reasons as well as other questions raised by stakeholders, we ultimately did not pursue this strategy. Another possible approach included the identification of baseline-period usage characteristics that are predictive of COVID-related consumption changes. However, in assessing the correlation between COVID impacts and nearly 30 computed usage parameters at an individual-meter level, none showed sufficient promise to warrant further investigation. The full results of this analysis are presented in Chapter 6.

IV. Stratified Sampling

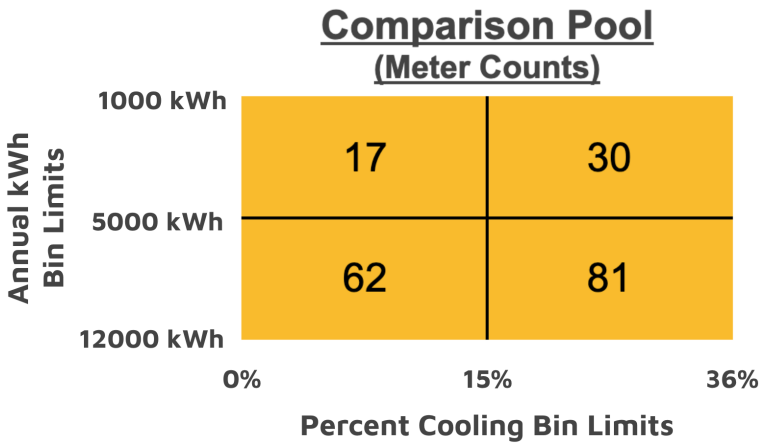
In stratified sampling, a comparison group is generated from a comparison pool by selectively eliminating members of the comparison pool until the remaining sample is representative of a treatment group. This elimination step is done on the basis of one or more characteristics that can be known or measured 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:

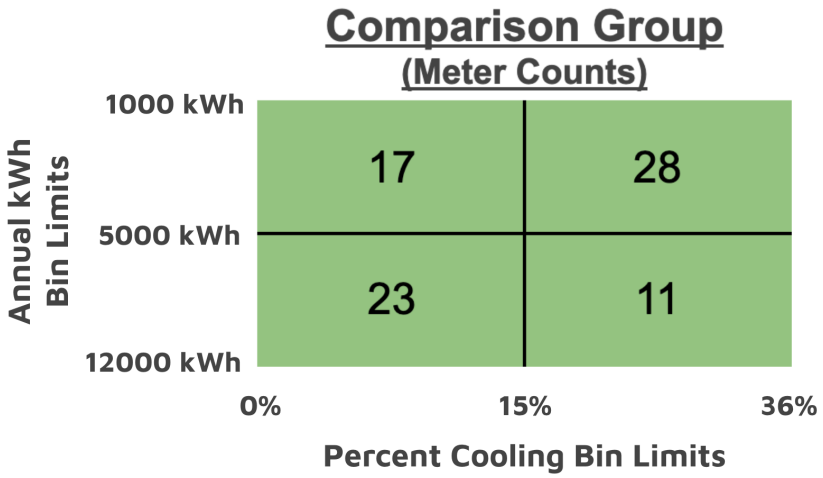
Comparison vs. Treatment
(Population Ratio)

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 first 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 the majority of 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.[7] 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.

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

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

minimum ratio of 4:1 is recommended for small Residential treatment groups (< 750 meters).[8] 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 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:

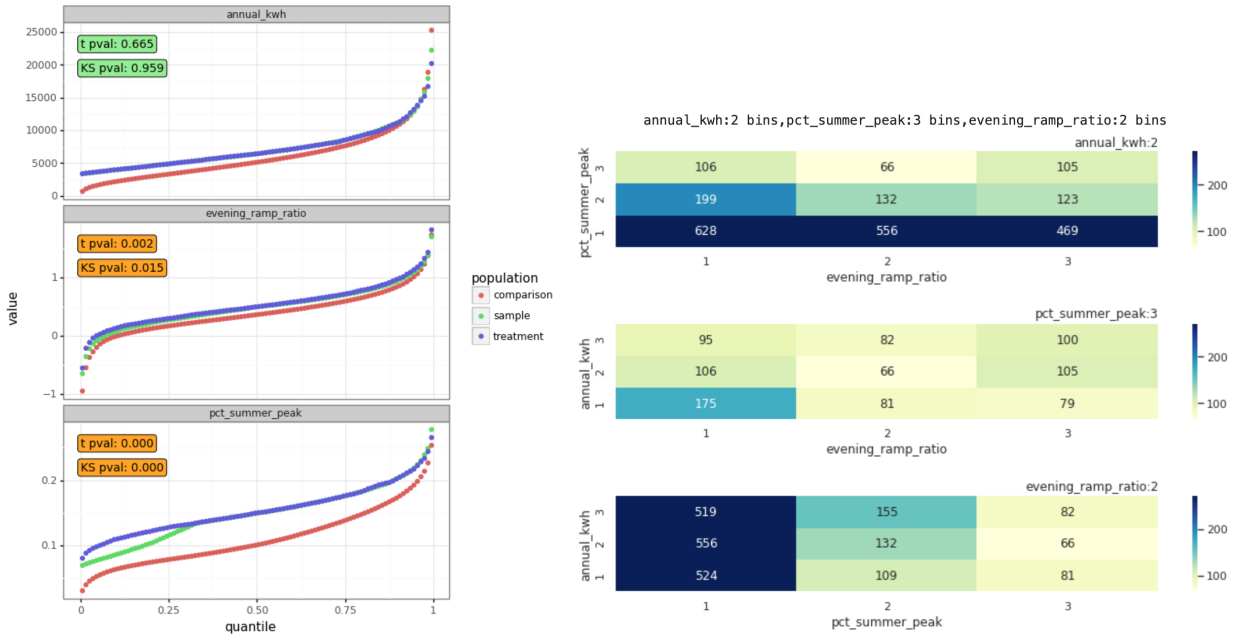
1. Customers were selected who were in the top 75% of total baseline period usage and the top 40% of their utility cost per MWh ratio. The latter metric Recurve calculates based on customers' avoided cost profiles using the California Public Utility Commission's 2020 avoided cost data.[9] Out of the initial 50,000-meter sample, 16,606 meters met these thresholds.
2. Of these 16,606 meters a random sample of 2,000 meters was selected.

[8] 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.
[9] https://www.cpuc.ca.gov/General.aspx?id=5267

3. The second 50,000-meter sample (see Chapter 2) was utilized as the comparison pool.
4. 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[10] and evening ramp ratio.[11]
5. For this exercise, maximum bins for each stratification parameter were set to 3, though this limit could be made significantly higher if beneficial.
6. 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 10 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.



[10] Defined as the percentage of a customer's total annual kWh usage that occurs from 4 - 9 pm during the months of June - September.
[11] Defined as a customer's average usage during hour 18 minus average usage during hour 14 divided by the total annual usage.

Figure 10: 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 10 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 11 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.

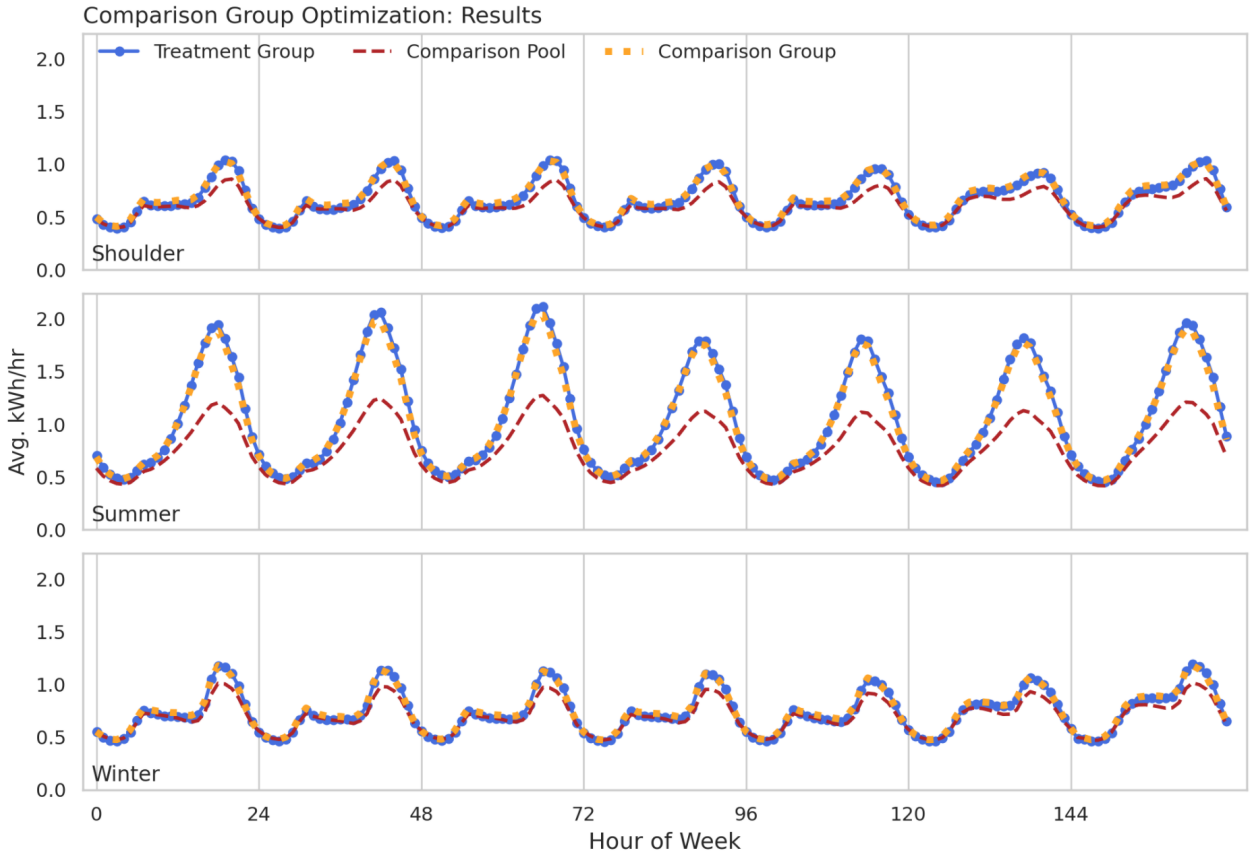


Figure 11: Baseline seasonal weekly load shapes for the treatment group, comparison pool, and comparison group resulting from the advanced stratified sampling scheme described in this chapter.

Just as important a test is how the comparison group behaves during the counterfactual period relative to the treatment group. Figure 12 shows that the good load shape match observed in the baseline period continues into the reporting period (the COVID period).

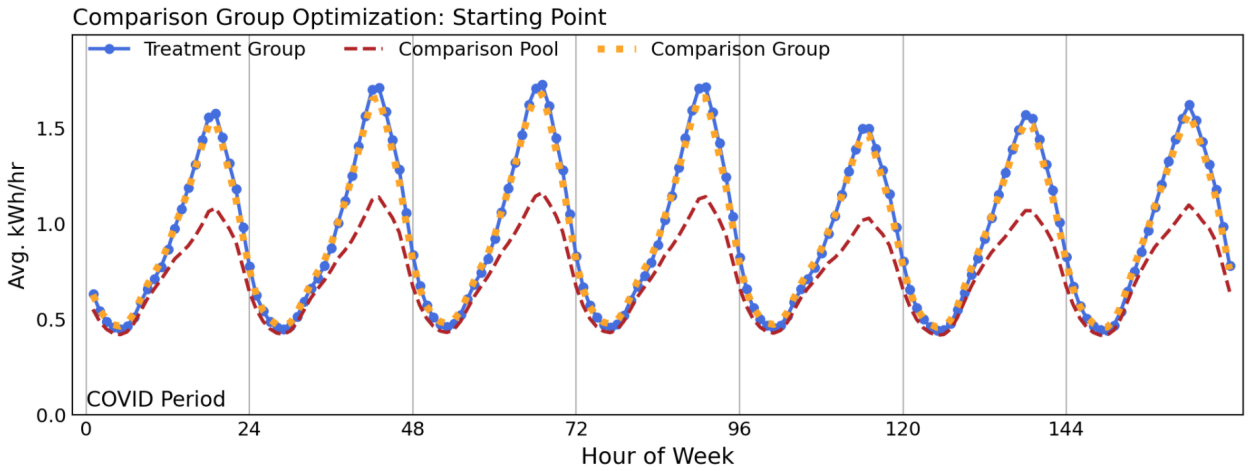


Figure 12: Reporting period (COVID-period) weekly load shapes for the treatment group, comparison pool, and comparison group resulting from the advanced stratified sampling scheme described in this chapter.

V. Site-Based Matching

In the site-based sampling method each treatment meter is matched with meters from the comparison pool to which it is most similar. Site based matching was initially thought to be too computationally intensive to be feasible at scale, but additional research and tools development in the open source community have shown not only its feasibility but its high rate of accuracy. Site-based matching is constrained by the size of the comparison pool, and usually requires a comparison pool to be significantly larger than the treatment group for reasonable results.

In site-based matching, a computation is performed on every combination of treatment and comparison pool meters to determine the similarity between each pair.[12] For an individual treatment/comparison meter pair this is done by calculating the distance between each feature data point. A feature may be an individual month's consumption, or a usage characteristic like the targeting features described above in the description of stratified sampling. For example, with monthly data the distance calculation can be performed by calculating the sum of squares across the monthly consumption values.[13] If using consumption data as features, the limiting granularity of the site-based matching calculation depends on the granularity of usage data. If hourly data are available then the distance calculation could be done on an 8760 basis. However this level of granularity may not be computationally feasible or necessary. In the case where hourly data are available, we recommend performing site-based matching on the basis of the average seasonal weekly load shapes as also described above for advanced stratified sampling.

In some cases it may be paramount to ensure the comparison group is a particularly good representation of the treatment group across specific features. For instance, in an air-conditioning program, matching based on cooling energy consumption as a primary feature may be justified. In a demand response program, features corresponding to peak-period usage could be prioritized. Therefore in site-based matching it is useful to be able to apply weightings to particular features. The GRIDmeter open source

[12] Or site aggregation

[13] 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.

sampling code that Recurve is releasing as part of this effort enables the user to weight features. GRIDmeter also allows the user to make comparison pool meters that have already been assigned to a treatment meter eligible for assignment to another treatment group meter. However, for most cases this is not recommended.

After performing these computations, each meter is matched based on a minimum calculated distance value. Matched meters are selected as part of the comparison group. 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 user also has the option to set a max tolerable distance threshold. When exceeding this threshold, a new meter will not be added to the group, guaranteeing a minimum threshold for similarity between the treatment and comparison group. 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.

In some cases, site-based matching has been found to be more accurate than stratified sampling when high granularity data are available. Generally, if hourly data are available with a large comparison pool, site-based matching can produce a more similar comparison group than stratified sampling.

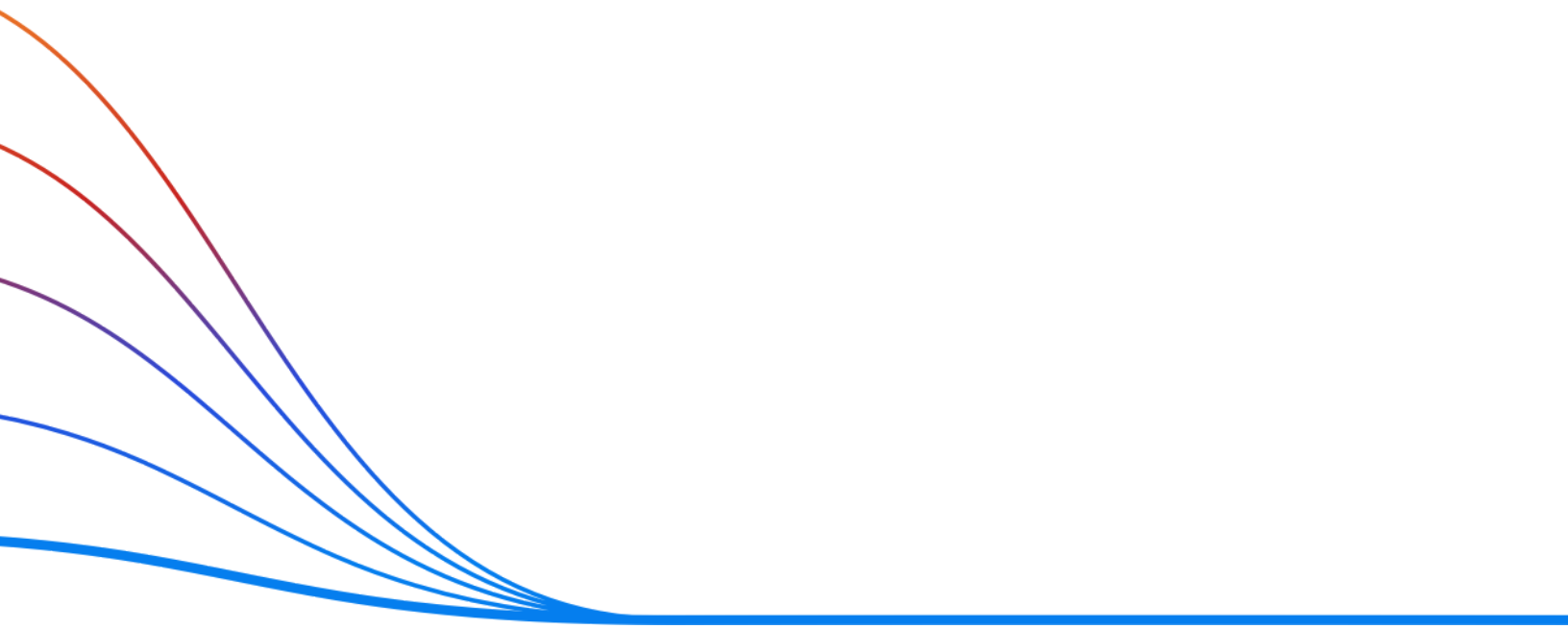
VI. Open Source GRIDmeter Codebase

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.[14]

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

RECURVE
SHAPE THE FUTURE OF ENERGY

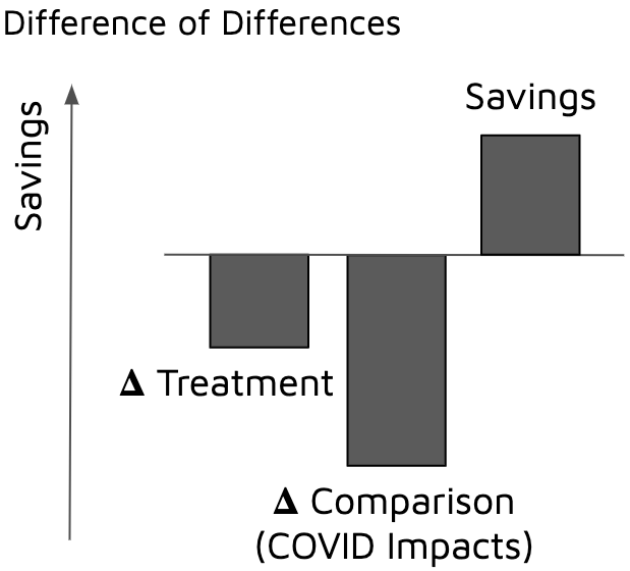
Chapter 4: Performing Difference of Differences Calculations



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 provides recommendations to specify 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:

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

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.

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

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.

$$Savings_{(Difference\ of\ Differences)} = Difference_{Treatment} - Difference_{Comparison}$$
$$= (Counterfactual_{Treatment} - Observed_{Treatment}) - (Counterfactual_{Comparison} - Observed_{Comparison})$$

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 today, 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 13 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.

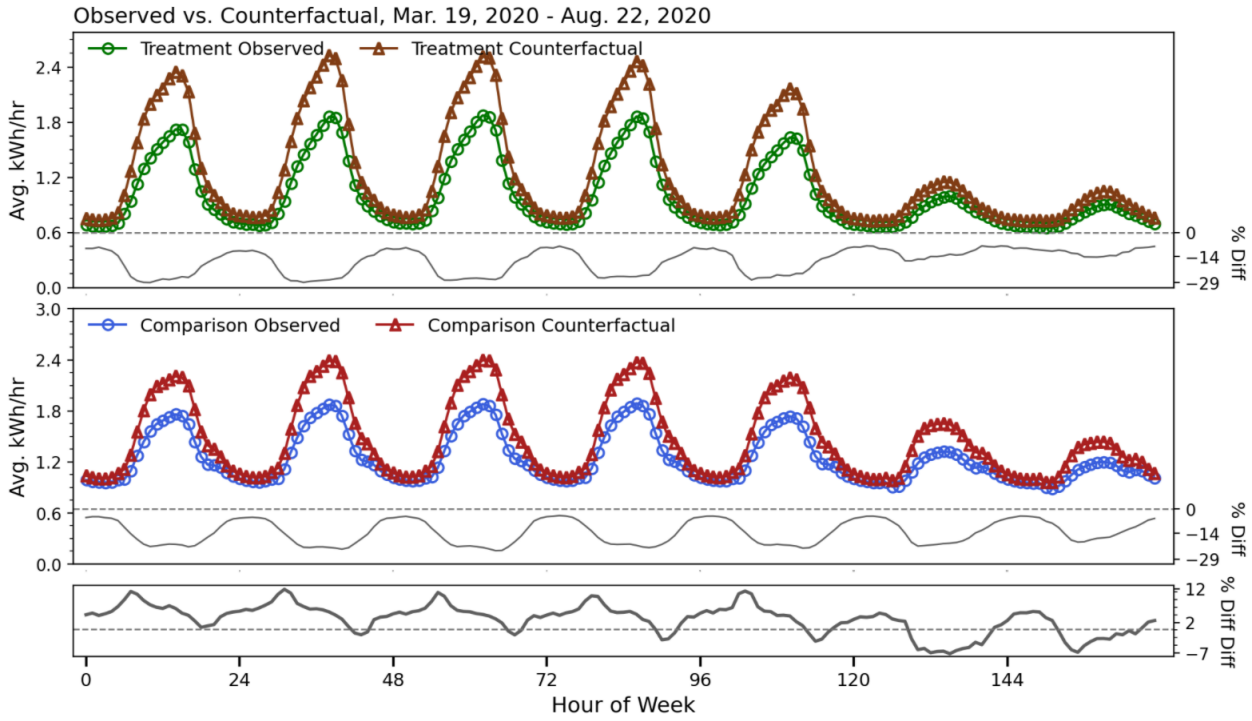


Figure 13: Average weekly load shapes and percent difference traces for each element of a difference of differences calculation for hypothetical treatment and comparison groups

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?
- 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) and this choice is therefore most appropriate for the long term.

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. 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 week, month, or quarter of a year may be too expensive and impractical. Therefore, as a middle ground, we recommend that at a minimum computing the savings of a single comparison group should be done at multiple points throughout the year to best capture the appropriate timelines for a program. This can be done by producing comparison group savings calculations where the reporting period is set to begin for each month of the year. The savings from these monthly "vintages" of a single comparison group can then be used to adjust treatment group savings via the difference of differences calculation for each monthly cohort of treatment group meters.

The right balance must be struck in the execution of comparison group vintages between temporal granularity and computational cost. Especially when performing hourly calculations on upwards of several thousand comparison group meters, CPU and cloud computing costs as well as the data infrastructure needed to reliably organize and utilize outputs in a transparent manner can become barriers. Along these lines, while the savings calculation is worth the effort to create vintages for the comparison group, for large comparison pools, it is not likely worth the cost to compute all possible stratification parameters, which themselves are derivatives of the baseline period calculation, for every possible baseline across all 12 months of the year. Though some meters may be lost due to data sufficiency requirements from the comparison pool by taking this approach, a reasonably-sized comparison pool should still have a sufficient number of meters available. However, treatment meters should not be discarded for not having a full baseline period available in the year leading up to program launch. For example, a treatment customer who participated in October of 2020 should not be required to have a full 12 months of data from Jan. 1 through Dec. 31 of 2019 to be included in the savings measurement. With these practicalities in mind, we provide the following stepwise approach to implement comparison group vintages for the difference of differences calculation:

a. If using stratified sampling or proportional sampling based on usage characteristics as described in Chapter 3, complete steps i and v. If using random sampling or proportional sampling based on geography alone, steps i and v can be skipped.

   i. Using the 12 months preceding the program year as a baseline period, compute stratification parameters for the entire comparison pool.

   ii. Compute monthly, daily or hourly savings depending on the granularity of consumption data for all meters in the treatment group per CalTRACK 2.0 methods.

   iii. Assign each treatment meter to a monthly cohort according to its program participation end date.

   iv. Compute percent savings for each treatment group monthly cohort:

$$\%\,Diff_{Treatment,i} = \sum(Counterfactual_{Treatment,i} - Observed_{Treatment,i})/\sum(Counterfactual_{Treatment,i})$$

where the summations are computed over all treatment group meters. This is done on an hourly, daily, or monthly basis depending on the granularity of the consumption data. The monthly cohort is determined by the program intervention end date.

v. With the same baseline periods utilized for step ii, compute stratification parameters for each member of the treatment group. (Note that several important stratification parameters, including heating and cooling loads are derived from outputs of CalTRACK calculations.)

vi. Complete comparison group sampling as detailed in Chapter 3.

vii. Compute savings using CalTRACK 2.0 methods for the comparison group for each monthly vintage. The baseline period for the first vintage will consist of the 365 days leading up to the first month of the program year. The reporting period for the first vintage will begin on the first day of the month beginning the program year and run for the subsequent 365 days. For instance, for a program year beginning in Jan. 2021, compute comparison group savings using CalTRACK 2.0 methods with a baseline period of Jan. 1, 2019 - Dec. 31, 2019 and a reporting period of Jan. 1, 2020 - Dec. 31, 2020. The second vintage will be shifted exactly 1 month forward for both baseline and reporting periods. This step is complete when comparison group savings are computed for all 12 monthly vintages.

viii. Compute savings for all comparison group monthly vintages:

$$\%\,Diff_{Comparison,i} = \sum(Counterfactual_{Comparison,i} - Observed_{Comparison,i})/\sum(Counterfactual_{Comparis}$$

where i indicates the monthly vintage and the summations are computed over all comparison group meters. This is done on an hourly, daily, or monthly basis depending on the granularity of the consumption data.

ix. For each monthly cohort compute the percent difference of differences:

$$\%\,Diff\ of\ Diff_{i} = \%\,Diff_{Treatment,i} - \%\,Diff_{Comparison,i}$$

This is done on an hourly, daily, or monthly basis depending on the granularity of the consumption data.

x. Compute monthly cohort savings by multiplying the percent difference of differences by the treatment group counterfactual:

$$Savings = \%\,Diff\ of\ Diff_{i} \times Counterfactual_{Treatment,i}$$

This is done on an hourly, daily, or monthly basis depending on the granularity of the consumption data.

xi. Compute total hourly, daily, monthly, or annual savings by summing results from all monthly cohorts as needed.

Best practice is to calculate savings on the most granular level possible. For example, if hourly data is available, savings should be calculated on an hourly percentage basis, and then summed.

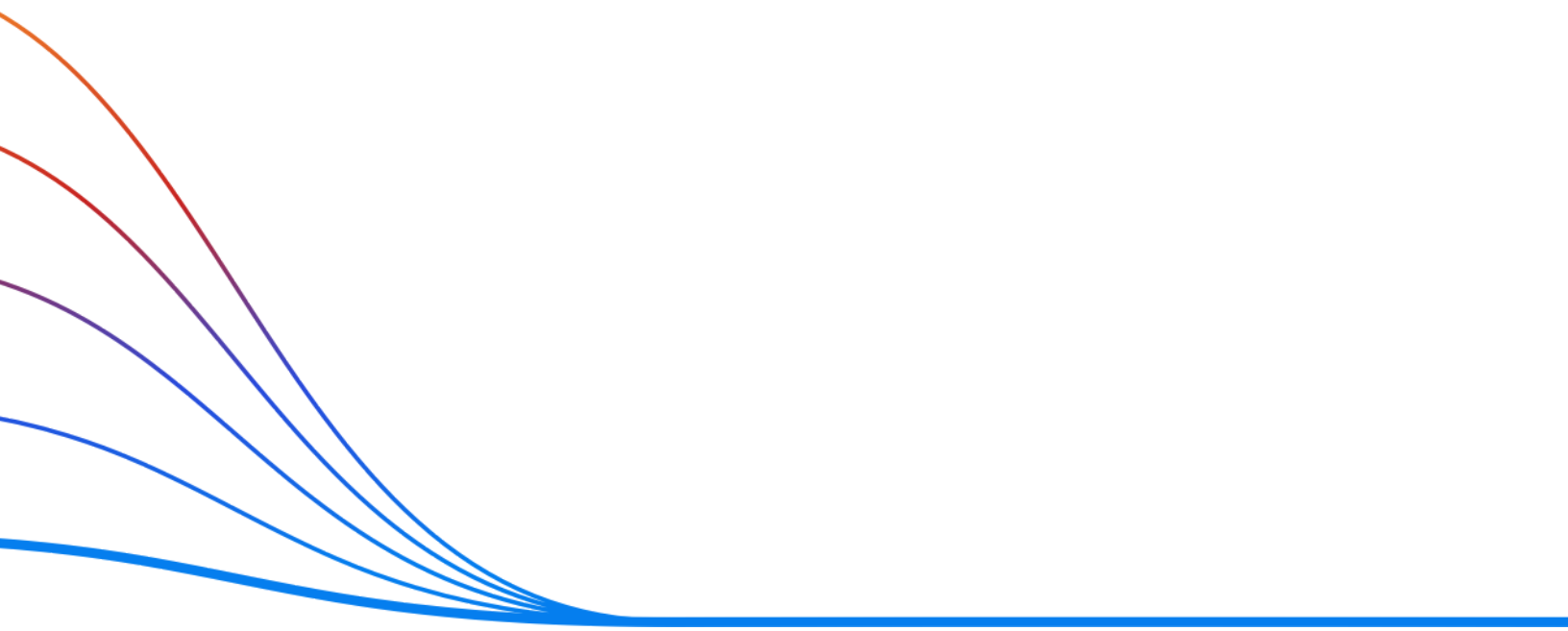
D. A Note on Computational Granularity

Because a percentage difference of differences calculation is a non linear 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.

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.[15] 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 14 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).

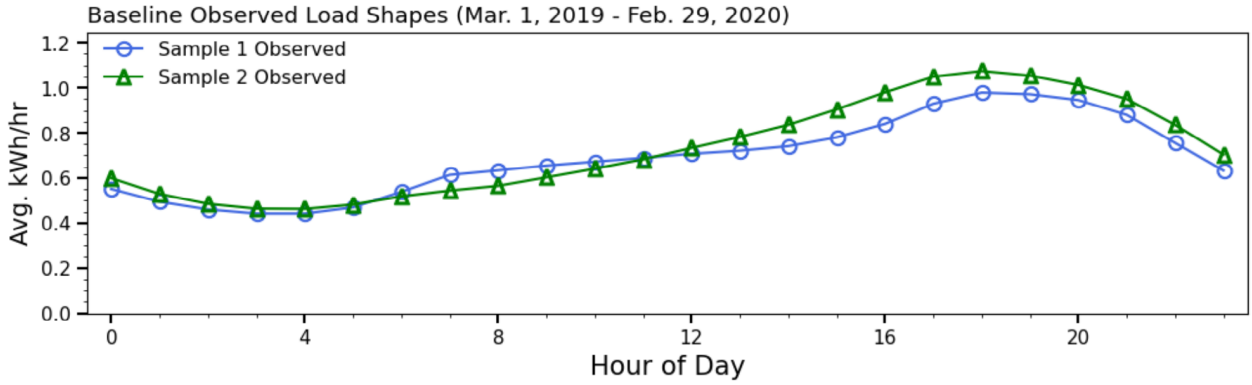


Figure 14: Average daily load shape of a residential non-solar MCE meter (blue circles) and a residential non-solar MCE meter in Pittsburgh (green triangles).

[15] 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 2 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 2

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 3 details the selection schemes explored here.

Table 3

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 15 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).

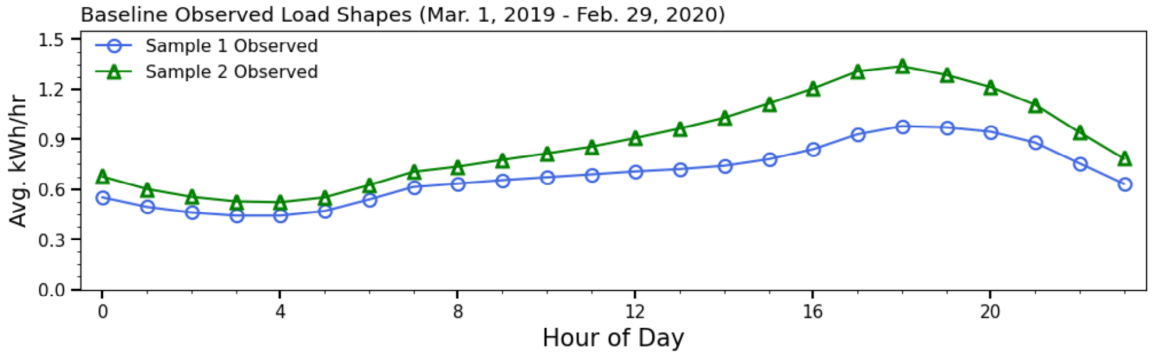


Figure 15: Average load shape of a residential non-solar MCE meter (blue circles) and a residential non-solar MCE meter in the top 75% and top 40% of all residential meters in annual usage and the percent of usage from cooling.

Table 4 provides a summary of results for these comparison group tests.

Table 4

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 4. 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 CalTRACK 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:

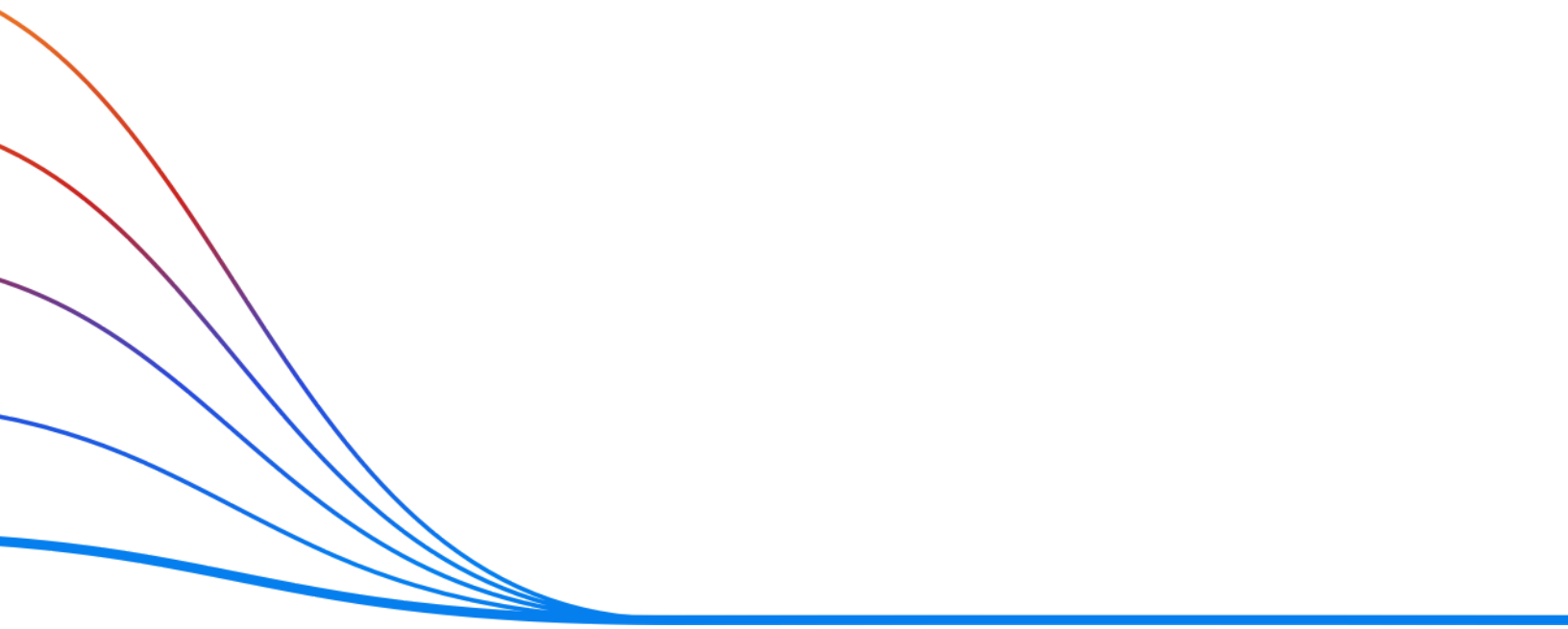
a. Were solar customers (identified by rate code or the presence of negative meter readings).
b. Had a total annual consumption in the baseline period greater than 50 MWh.
c. Had fewer than 329 days with at least one meter reading in the baseline period.
d. Had more than 15% of hours with null meter 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 5:

Table 5

	% 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 CalTRACK savings calculation. With the exception of 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 6.

Table 6

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 CalTRACK 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 7 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 7

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.