Technical Report

Comparison Group Identification for Impact Evaluation

Prepared for Energy Trust of Oregon

Hassan Shaban, Ph.D.
Senior Data Scientist
Open Energy Efficiency
hassan@openee.io

McGee Young, Ph.D.
Chief Product Officer
Open Energy Efficiency
mcgeeyoung@openee.io

Table of Contents

Executive Summary

Energy Trust of Oregon and Open Energy Efficiency conducted a study to test different methods of identifying comparison groups for impact evaluations. The findings of the study were intended to inform the implementation of standardized, automated impact evaluations in Energy Trust's Automated Meter Data Analytics Platform.

Use case

There is a need to increase the speed and efficiency of conducting billing analyses and a desire to implement more standardized methods, in order to provide consistent and faster feedback to energy efficiency program managers and third party implementers. These automated approaches can be applied more efficiently and consistently than standard EM&V practice, enabling utilities and markets to optimize solutions and programs, and support private investment and risk management.

Main findings

Several methods of comparison group identification, as well as several methodological issues were investigated as part of this study. These methods were evaluated using out-of-sample testing as well as using a number of equivalence metrics. There were some differences in the mean savings estimates of different methods, however in many cases, the uncertainty bounds of the different methods overlapped with each other, indicating that these differences were not always statistically significant. It is unclear if there is one "best" method - in particular, monthly consumption matching and future participant groups offered similar levels of performance for different datasets.

Recommendations

The primary recommendation when implementing automated comparison group identification is to automate the calculation, not the interpretation of results. This can be applied by using several different methods simultaneously and several quality metrics to judge the appropriateness of a comparison group. Three methods were recommended in particular (depending on data availability): individual customer matching on monthly consumption, stratified sampling of future participant groups and stratified sampling of past participant groups. This holistic approach would work well for impact evaluations, however, if comparison groups were to be factored into payments in pay-for-performance settings, then we recommend that the comparison group identification method be contractually set before the launch of a pay-for-performance procurement and accommodated in the program design.

Overall, this study has shown that automated data-driven methods can produce comparison groups quickly and consistently, and can support a range of use cases. Further work is planned to continuously improve these recommendations as they are applied with more diverse datasets.

MEMO

Date: October 26, 2018
To: Board of Directors
From: Dan Rubado, Evaluation Project Manager
Subject: Staff Response to the Open EE Technical Report on Comparison Group Identification

Energy Trust contracted with Open EE to build an automated, web-based tool to conduct impact analysis of residential efficiency measures based on utility billing data. Open EE uses industry-standard methods, similar to the Princeton Score-keeping Method, to weather-normalize energy usage data and conduct pre/post analysis. In addition to weather normalization and pre/post analysis, impact analysis requires a quasi-experimental design, in which a comparison group that resembles the treatment group is selected. The comparison group represents the "counterfactual" and helps answer the question of what would have happened to energy usage in the treatment group in the absence of an intervention. Comparison groups help control for the effects of atypical weather and exogenous trends in energy usage.

There are many methods for selecting comparison groups for quasi-experimental studies, but no agreed upon best practices. This report quantitatively compares several commonly used methods to assess how well they perform, in terms of representing the treatment group and providing an unbiased counterfactual case. It also explores several other analytical issues that are important to impact analysis. The report documents the analysis methods employed by Open EE, makes recommendations about which comparison group and analysis methods to use for impact analysis of residential energy efficiency measures going forward, and how to monitor the performance of those methods.

Unfortunately, Open EE was unable to conclusively identify a "best" method for selecting a comparison group for residential impact analysis. However, monthly consumption matching and future participant groups appeared to perform similarly well across a variety of metrics and were recommended above more simplistic techniques. Open EE recommends using several different comparison group methods, then comparing and combining estimates, which may provide more stable results than a single method. They also recommend continuing to monitor the quality of matches and performance in the baseline period for each comparison group method. They have several additional recommendations on specific analytical issues that we generally agree with.

The impact analysis tool that Open EE is building for Energy Trust will incorporate all the recommended analytical and comparison group methods. Once completed, Energy Trust's evaluation team will have the capability to conduct utility billing analyses of residential efficiency measures much more quickly than in the past. Other benefits will include more standardized analysis methods, less staff time required for analysis, and lower costs per measure analyzed. As a result, the evaluation team intends to substantially increase the volume and frequency of residential measures that we analyze. In addition, the methods developed through this work will also be leveraged in Energy Trust's Residential Pay for Performance Pilot, launching in 2019, for which Open EE will be quantifying the energy savings.

1. Introduction

Energy Trust of Oregon (“Energy Trust”), in collaboration with Open Energy Efficiency (“OpenEE”) is seeking to implement functionality that facilitates standardized impact evaluations in their Automated Meter Data Analytics (“AMDA”) Platform. Energy Trust aims to implement this new functionality for use in analyzing the impacts of residential energy efficiency interventions, with the ultimate goal of increasing the speed and efficiency of conducting billing analyses using standardized methods, in order to provide better and faster feedback to program managers and third party implementers.

The AMDA Platform is currently based on the open source OpenEEmeter, which implements methods developed through the CalTRACK process. These methods undergo continuous testing and improvement through an annually-convened technical working group to ensure that the results from this platform are as robust and reliable as possible. CalTRACK is an open-source set of methods that focus specifically on calculating site-based, weather-normalized metered energy savings for determining payments under pay-for-performance programs. CalTRACK methods describe how to calculate whole-building site-based savings that result from any mix of measures, building types, and consumer behavior, but not the amount of savings that can be attributed to any particular measure. CalTRACK includes methods for handling billing/monthly data as well as daily and hourly data. Additionally, CalTRACK includes guidance on aggregating individual site-level projects to portfolios (groupings of similar energy efficiency projects, whose savings estimates are more robust than individual projects).

The result of the CalTRACK 1.0 process (2017) was an initial set of technical requirements and methods for calculating and reporting normalized metered savings for residential energy efficiency projects in California based on standard input data formats and analysis methods. The CalTRACK 2.0 working group (2018), comprising utility representatives, regulators, and building energy experts, investigated certain limitations in the original methods, thereby making the methods applicable to a wider range of building types and climate zones, in addition to adding methods for qualifying buildings for pay-for-performance and for estimating portfolio loadshapes.

Weather normalization of billing data in CalTRACK follows certain model foundations in literature (PRISM, ASHRAE Guideline 14, IPMVP Option C and the Uniform Methods Project for Whole Home Building Analysis[1]). Building energy use is modeled as a combination of base load, heating load, and cooling load. Heating load and cooling load are assumed to have a linear relationship with heating and cooling demand, as approximated by heating and cooling degree days, beyond particular heating and cooling balance points as shown in Figure 1. A number of candidate OLS models are fit to the consumption data using different combinations of heating and cooling balance points (ranging from 30 to 90 F) and different sets of independent variables (Figure 1). The

[1] https://www.nrel.gov/docs/fy17osti/68564.pdf

model with the highest adjusted R-squared that contains strictly positive coefficients is selected as the final model and used to calculate normalized energy usage. Further details about these methods are housed at the CalTRACK website[2] and the underlying data science can be found on Github[3].

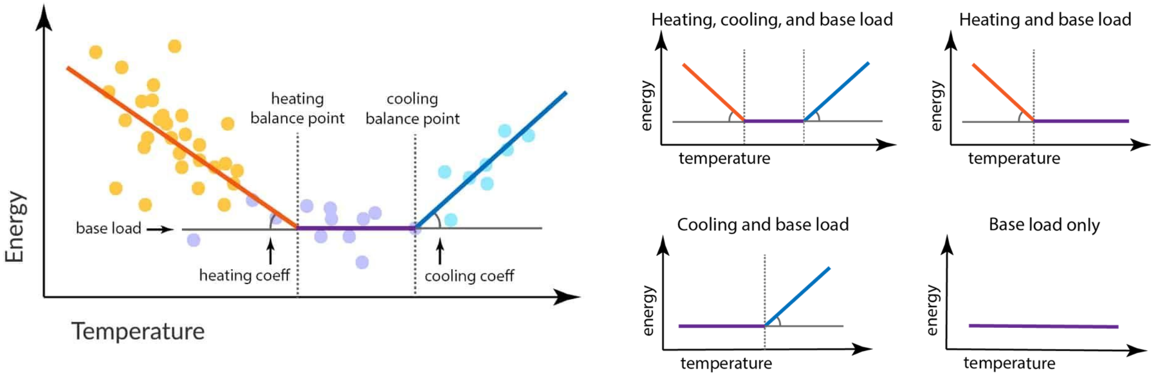


Figure 1. CalTRACK model nomenclature (left) and types of CalTRACK candidate models (right).

In addition to standard program impact evaluations, and with the imminent launch of pay-for-performance programs in Oregon, there is a need to better understand the pros and cons of different comparison group identification methods for pay-for-performance program designs. In contrast to standard deemed or custom programs, in a pay-for-performance design, program administrators offer incentives or other payment either directly to customers or to aggregators (entities that implement or procure energy efficiency in portfolios of buildings) in exchange for energy savings that are typically measured during an agreed-upon performance period. This makes it necessary for the payment calculation methodology (including comparison group identification) to be (a) fully transparent to both parties and (b) contractually defined prior to the launch of a program.

The OpenEEmeter implementation defines two methods of calculating metered savings (also referred to as "changes in consumption"):

a. *Payable savings*: One model is fit to the baseline (pre-intervention) period consumption. A counterfactual is calculated by applying reporting (post-intervention) period weather data to the baseline model. Savings are calculated as the difference between this counterfactual and the actual energy consumption during the reporting period. This is the default method defined in CalTRACK for pay-for-performance programs.

b. *Normal Year savings*: Two models are fit - one for the baseline (pre-intervention) period and one for the reporting (post-intervention) period. Typical Meteorological Year weather is applied to both models to calculate the Normalized

[2] https://docs.caltrack.org
[3] https://github.com/CalTRACK-2/caltrack

Annual Consumption (NAC) for both periods, and savings are determined by subtracting the two NACs. This method is useful for estimating long term impacts of efficiency programs and this is the method that has been used to quantify savings in the current study.

These two types of savings estimates do not include the effect of unmeasured factors on energy consumption. Quasi-experimental approaches offer one way to control the effect of exogenous trends in energy use on savings estimates. For example, using a two-stage approach[4], site-level models may be fit to treatment and comparison group consumption followed by a difference-of-differences calculation to estimate the savings net of exogenous trends and other market effects, including savings from codes and standards, midstream and upstream programs and natural adoption of energy efficiency. These approaches may not completely eliminate biases in savings due to self-selection, however, they may be the only feasible option for newer programs that do not have mature participation cohorts for other comparison group specifications (e.g. past/future participants).

The main objectives of the present study are as follows:

a. Implement a set of standard comparison group identification methods that are amenable to automation in Energy Trust's AMDA platform.
b. Test the implemented methods with two types of residential measures.
c. Identify any user-defined parameters or methodological choices to which the results may be sensitive and recommend default values.
d. Prepare recommendations for automating comparison group identification for impact evaluations of standard programs and pay-for-performance programs.
e. Prepare a technical report for review by Energy Trust's panel of outside expert billing analysis reviewers and Board Evaluation Committee.

This report summarizes the results of the study and presents an initial set of recommendations. It is organized as follows:

- Section 2 begins with a description of the test dataset and methodologies used in the present study. The different comparison group matching methods are then described and their results are discussed.
- Section 3 describes the tests that were performed to isolate the effects of a number of methodological choices on comparison group matching results.
- Section 4 summarizes the results and presents the savings calculated using various comparison group matching methods.
- Section 5 includes recommendations for implementing automated comparison group identification for different use cases.

[4] Agnew, K.; Goldberg, M. (2017).Chapter 8:Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures. Golden, CO; National Renewable Energy Laboratory.NREL/SR-7A40-68564. http://www.nrel.gov/docs/fy17osti/68564.pdf

2. Comparison group identification methods

2.1 Test data

Energy Trust provided billing data for all residential customers in Oregon, spanning 2011 through 2018. In addition, program participation data was provided by Energy Trust's third-party implementer. Two measures were selected for this initial analysis: smart thermostats and ceiling insulation. However, it was decided to focus on the gas savings of ceiling insulation participants, by virtue of its simplicity, larger group sizes and the availability of a recent comprehensive EM&V report for comparison.

Most of the following results are provided for gas customers who installed the ceiling insulation measure in the 2014 program year. Results for other years and for smart thermostats were also analyzed and did not show remarkable differences from this subset; these results are included in the Appendix for reference. *Portfolio-level methods* generate comparison groups without considering the features of individual customers, in contrast to *individual matching methods* which identify specific comparison customers for each treatment customer, by matching on location and consumption patterns. No data other than consumption and geographical location was used in any of the methods to ensure that they could be universally applied without depending on external data sources (e.g. demographic data or building characteristics).

There were approximately 600 program participants in the treatment group and a random sample of 200,000 non-participants was used as the pool of comparison group candidates (over 1.2 million were available, but a smaller pool was considered sufficient and significantly reduced the required computational resources for these tests). Consumption data preparation and cleaning followed best practices defined in the CalTRACK 2.0 billing methods.[5] Billing period consumption values were converted to usage per day values aligned with calendar months for subsequent analysis.

[5] The CalTRACK 2.0 methods can be accessed here: docs.caltrack.org.

Table 1. Summary of analyzed datasets.

Measure	Program Year	Fuel	Comparison group method(s)	Years of billing data analyzed	Number of customers	Relevant sections in the report
Ceiling Insulation	2014	Gas	All	2013-2015	601	2; 3
Ceiling Insulation	2014	Elec.	Future participants, Monthly consumption matching	2013-2015	771	A3
Ceiling Insulation	2015	Gas, Elec.	Future participants, Monthly consumption matching, Stratified sampling	2014-2016	637 (Gas) 750 (Elec.)	2.8; A4; A5
Ceiling Insulation	2013	Gas, Elec.	Future participants, Monthly consumption matching	2012-2014	648 (Gas) 769 (Elec.)	A1; A2
Smart thermostats	2015	Gas, Elec.	Future participants, Monthly consumption matching, Stratified sampling	2014-2016	434 (Gas) 425 (Elec.)	2.8; A6; A7

Analysis Periods

Different portions of the analysis used different time periods of consumption data, therefore, it is useful to clearly define these time periods and where they were used. Consider a project with an installation date on a particular day *d* in a particular month *m* in a particular program year *y*. The year before the program year is labelled as *y-1*, the year prior to that as *y-2* and so on, while the years following the program year are labelled *y+1*, *y+2* etc. (Figure 2) In all cases, the billing period that contains the project installation was dropped from the analysis. Other sections of the analysis use the following time periods, as shown in Table 2.

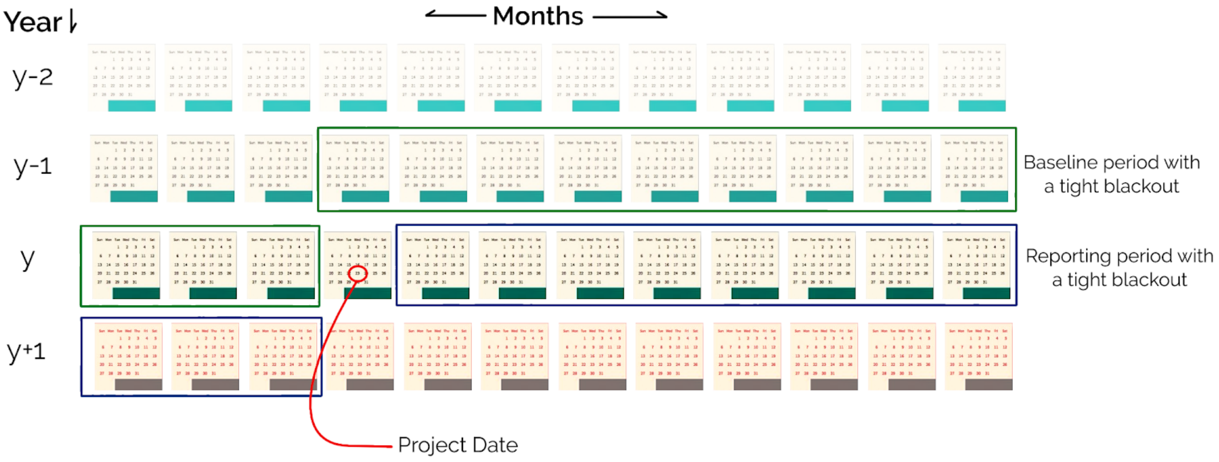

Year
Months
y-2
y-1
y
y+1
Baseline period with a tight blackout
Reporting period with a tight blackout
Project Date

Figure 2. Nomenclature for analysis time periods. Note that billing periods do not necessarily conform to calendar months

Table 2. Summary of analysis periods for different use cases.

Use case	**Group**	**Baseline period**	**Reporting period**
Consumption matching/ Group selection	Treatment group and individually matched comparison groups with tight blackout	12 months preceding the installation billing period (Fig. 2)	-
	Treatment group and individually matched comparison groups with loose blackout	Year *y-1*	-
	Groups identified using random and stratified sampling	Customers with sufficient data in years *y-1*, *y* and *y+1*	
	Future participant group	Participants from Year *y+1*	-
Equivalence tests	Treatment group and individually matched comparison groups with tight blackout	12 months preceding the installation billing period (Fig. 2)	-
	Treatment group and individually matched comparison groups with loose blackout; Groups identified using random and stratified sampling	Year *y-1*	-
	Future participant group	Year *y-1*	-
Savings estimates	Treatment group and individually matched comparison groups with tight blackout	12 months preceding the installation billing period (Fig. 2)	12 months following the installation billing period (Fig. 2)
	Treatment group and individually matched comparison groups with loose blackout; Groups identified using random and stratified sampling	Year *y-1*	Year *y+1*
	Future participant group	Year *y-1*	Year *y*
Out-of-sample testing	All	Year *y-2*	Year *y-1*

For example, for the participants who installed ceiling insulation in the 2014 program year, matching was performed using consumption data in 2013 with a loose blackout and using consumption data from the 12 months immediately preceding the intervention with a tight blackout; all equivalence metrics were calculated in 2013; savings calculations were done using 2013 and 2015 data for the random and stratified sampling methods, as well as when using a loose blackout; savings were estimated using 2013 and 2014 data for the 2015 future participant group (since they had project installations in 2015); and for the individually matched groups using a tight blackout, each comparison group customer was assigned the same blackout period as their corresponding treatment group match. Out-of-sample testing was performed using 2012 as the baseline and 2013 as the reporting period for all groups.

2.2 Equivalence metrics

Given two groups of customers, a treatment group comprising program participants and a comparison group comprising non-participants, several methods were used to evaluate the equivalence of their consumption in the portfolio baseline period (2013). The first three are visual, whereas the last three are numerical and could be used to automate the evaluation of comparison groups.

a. *Annual consumption histogram*: Allows the comparison of distributions of annual consumption for the two groups.

b. *Annual consumption Q-Q plot*: Allows more granular quantile-level comparison of the distributions of annual consumption.

c. *Plot of monthly energy consumption during the baseline year*: Allows comparison of aggregate monthly consumption in the portfolio baseline period.

d. *P-value from the t-test of annual consumption for the two groups*: Determines whether the means of the two groups are significantly different. A p-value larger than 0.05 indicates that there is no significant difference between the two group means.

e. *P-values from the t-tests of monthly consumption for the two groups*: Twelve p-values are calculated - one for each month in the portfolio baseline period. A p-value larger than 0.05 indicates that there is no significant difference between the two group means for a particular month.

f. *Kolmogorov-Smirnov (K-S) test for monthly consumption*: This test checks whether the distributions of monthly consumption are similar for the two groups. The K-S test yields a test statistic and a p-value for each month - a p-value larger than 0.05 indicates that there is no significant difference between the consumption distributions of the two groups for a particular month.

The various comparison group identification methods that were tested in the present study are discussed in the following sections. All of these methods were applied without a geographical screen (i.e. any customer within Oregon could be a potential match) and sampling was done without replacement.

2.3 Random sampling

This method involves selecting a random sample of non-participants (5x the number of participants) to serve as the comparison group. It is by far the simplest method of identifying a comparison group, but there is no guarantee that the treatment and comparison groups are similar in any way (except for the fact that they are in Oregon and use the same fuel). In the present study, this method is used as a naive model for comparison.

Figure 3a shows a comparison of the histograms of annual gas usage for the test dataset. In general, the matching appears to be reasonably close for such a basic method, indicating that the participant group is somewhat representative of the larger population. At closer inspection of the Q-Q plot in Figure 3b, it appears that the comparisons seem to have lower usage at the very low end and the very high end of the spectrum.

Figure 3c illustrates some obvious differences in consumption at the monthly level, with the comparisons having higher consumption in winter and lower consumption in summer, while Figure 3d shows that the differences in mean monthly consumption are significant ($p<0.05$). Except for the month of May, the distributions of monthly consumption are also significantly different as illustrated by the monthly K-S test.

This method is not recommended as it will likely bias gross savings results in unpredictable ways. Moreover, it is not a deterministic method and is highly sensitive to implementation details, e.g. the candidate pool and the random number seed that is used when sampling.

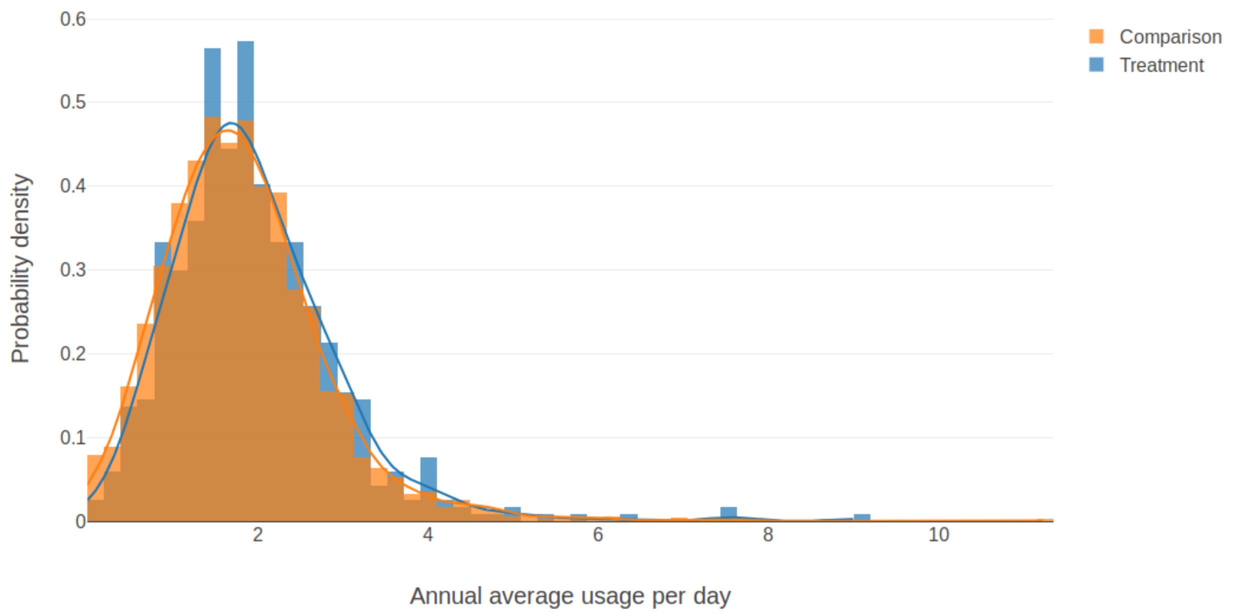


Figure 3a. Histograms of average annual gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined using random sampling ($n_{treatment}$ = 601; $n_{comparison}$ = 2478).

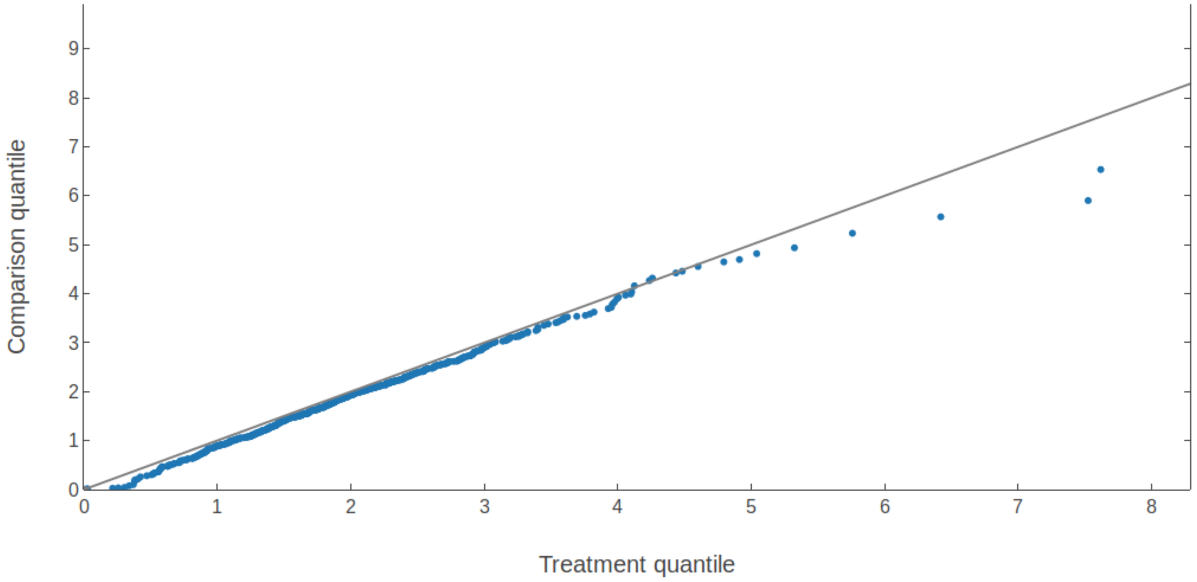


Figure 3b. Q-Q plot of annual average gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined using random sampling ($n_{treatment}$ = 601; $n_{comparison}$ = 2478).

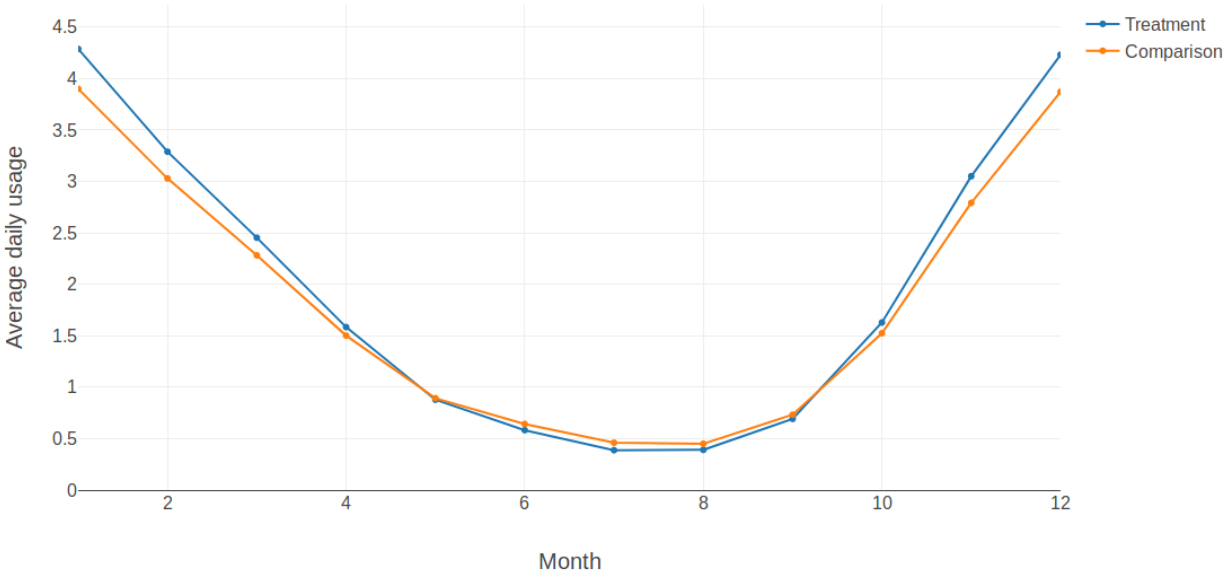


Figure 3c. Comparison of monthly average gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined using random sampling ($n_{treatment}$ = 601; $n_{comparison}$ = 2478).

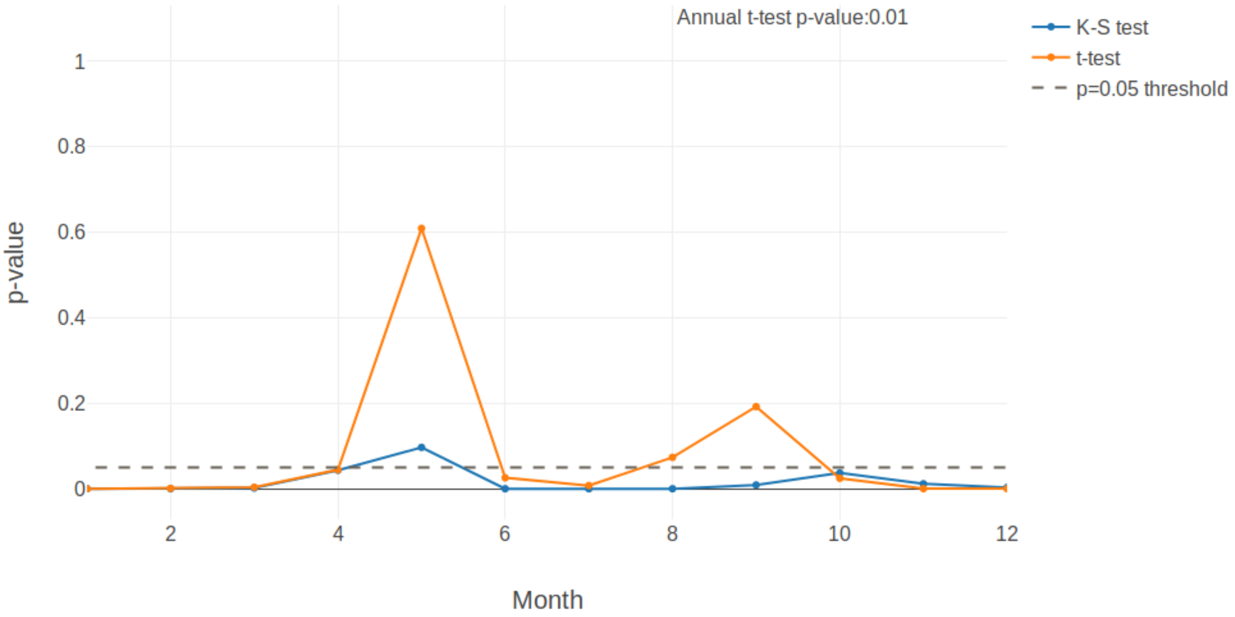


Figure 3d. Equivalence metrics for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined using random sampling. P-values larger than 0.05 indicate no significant differences in the mean (t-test) or distribution (K-S test) of monthly consumption ($n_{treatment}$ = 601; $n_{comparison}$ = 2478).

2.4 Future participants

At the other end of the spectrum, compared to purely random sampling, comparison groups comprising future participants are considered to be representative of participants in most aspects (observable and non-observable). For example, future participants are known to be eligible to receive the measure, and for some measures, they may have the same baseline equipment as the participants. Future participants have the same propensity to participate in the program as participants, thus reducing or eliminating self-selection bias, something that is otherwise difficult to control for in a quasi-experimental study. More comprehensive data is typically collected for future participants, allowing for potentially better matching and more insightful analysis.

From a practical perspective, future participant groups may be difficult to construct for all measures, unless a program has been running for multiple years and is considered stable with sufficient data collection over the analysis period. Sample sizes for the comparison group may also be constrained if using future participants.

The current set of future participants comprises all customers who installed the ceiling insulation measure in the 2015 program year. Figures 4a and 4b show that the annual gas consumption for the treatment group matched relatively well, except for a handful of customers with very large gas usage. The monthly consumption (Figure 4c) for the two groups also matched very well, with the treatment group having slightly higher consumption in all months. The mean annual consumption for the two groups were not significantly different (p=0.08) and Figure 4d shows that the distributions of monthly consumption were also very similar for all months (except for April, which has marginally significant difference).

These results are surprisingly good, considering that consumption was not taken into account at all during comparison group identification. The reasons for the slight positive bias in monthly consumption for the treatment vs. the comparison group may need further investigation.

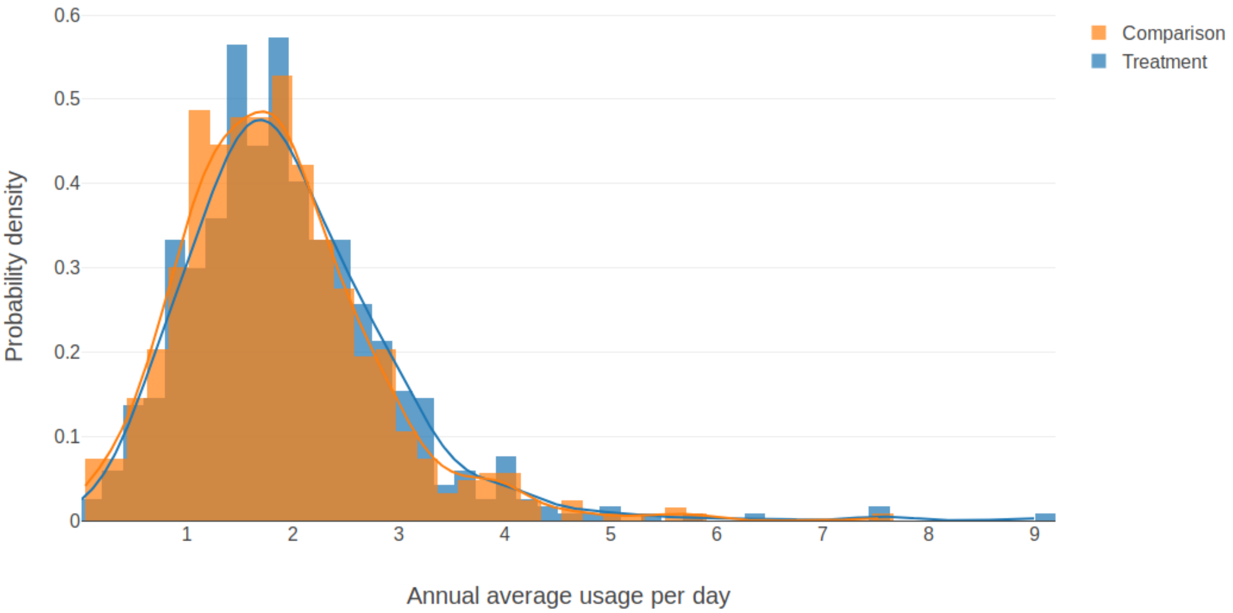


Figure 4a. Histograms of average annual gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a future participant comparison group (Ceiling insulation, 2015 program year) ($n_{treatment}$ = 601; $n_{comparison}$ =632).

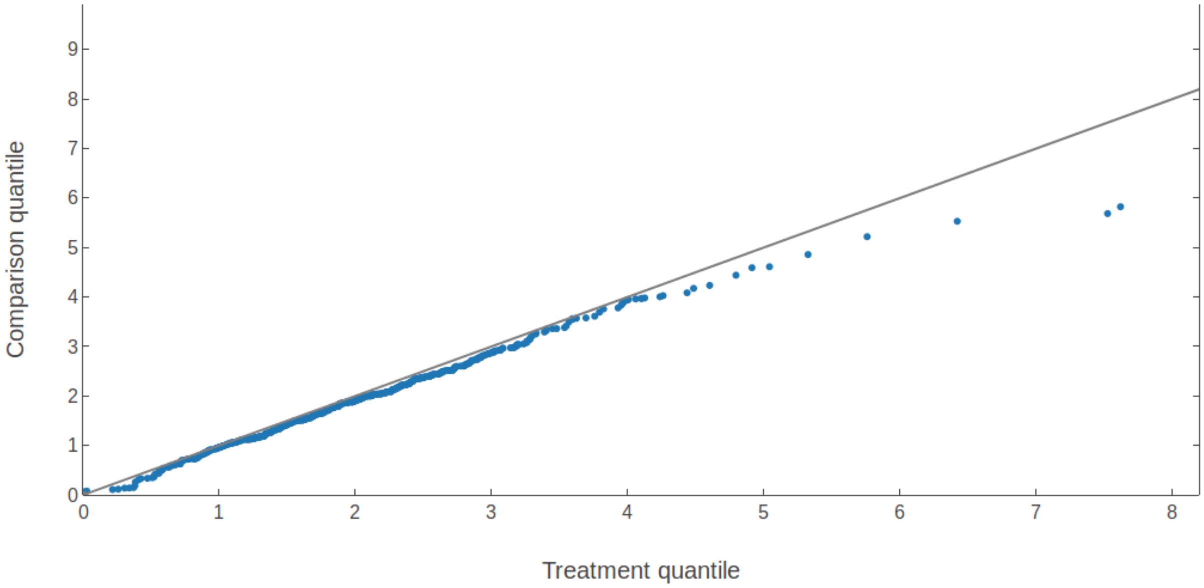


Figure 4b. Q-Q plot of annual average gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a future participant comparison group (Ceiling insulation, 2015 program year ($n_{treatment}$ = 601; $n_{comparison}$ =632)).

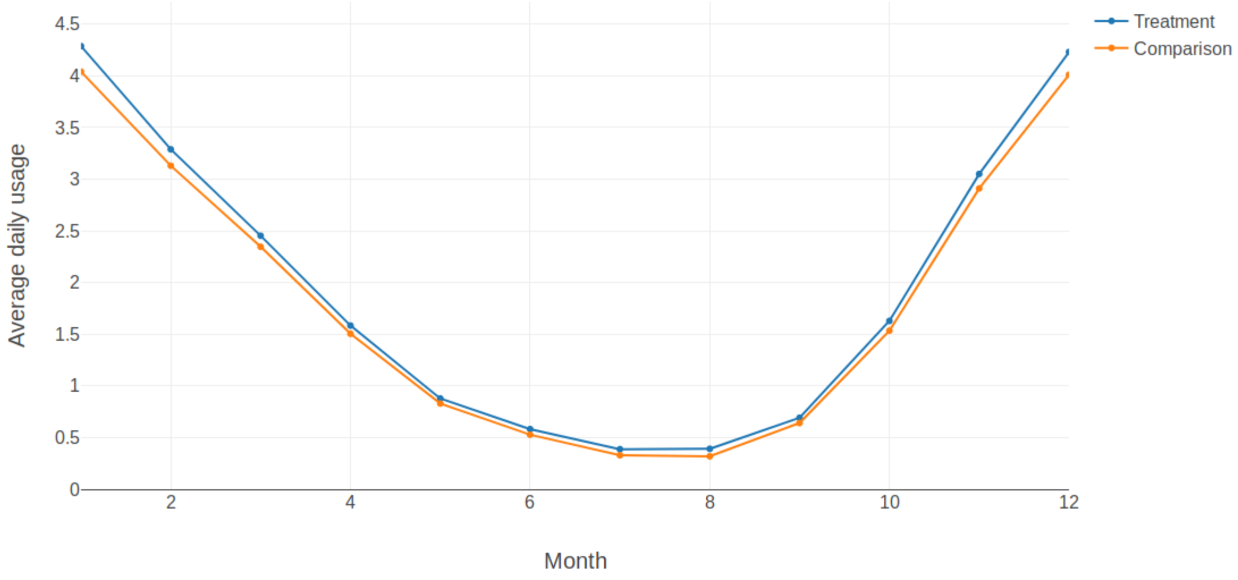


Figure 4c. Comparison of monthly average gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a future participant comparison group (Ceiling insulation, 2015 program year) ($n_{treatment}$ = 601; $n_{comparison}$ =632).

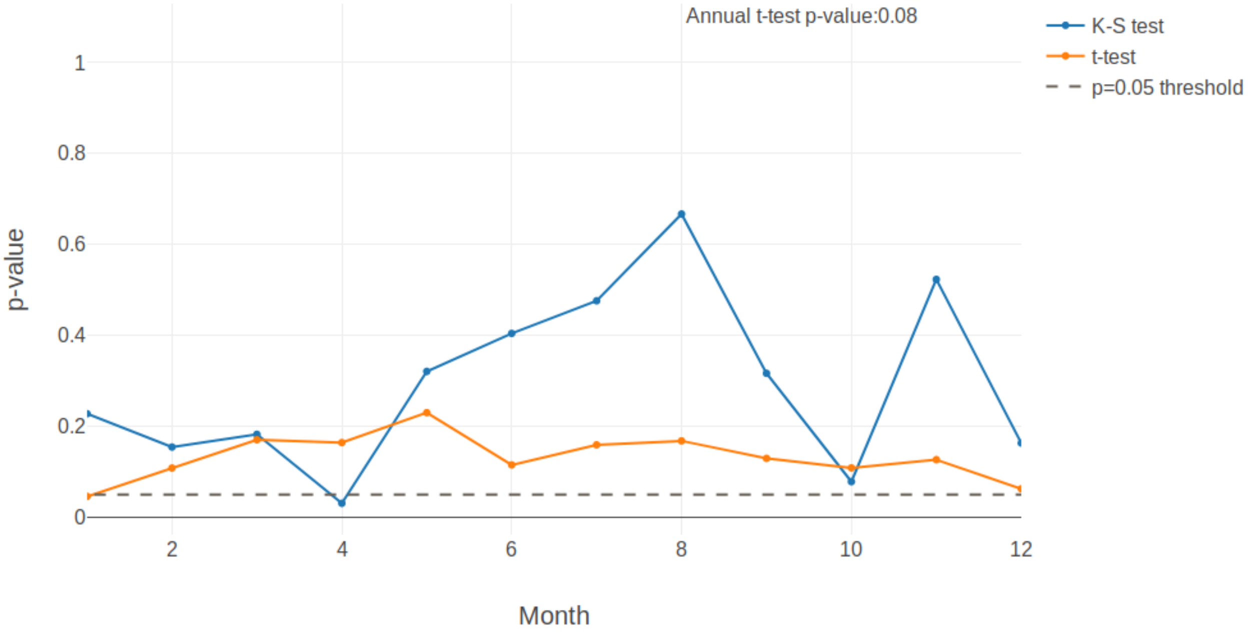


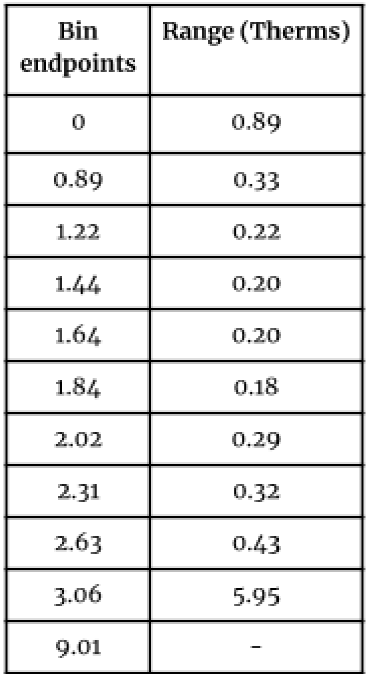
Figure 4d. Equivalence metrics for the treatment group (Ceiling insulation, 2014 program year) and a future participant comparison group (Ceiling insulation, 2015 program year). P-values larger than 0.05 indicate no significant differences in the mean (t-test) or distribution (K-S test) of monthly consumption ($n_{treatment}$ = 601; $n_{comparison}$ =632).

2.5 Stratified sampling

Stratified sampling was applied by first splitting customers into deciles, then selecting a random sample (5 times the number of treatment customers) within each corresponding bin in the comparison group pool of non-participants. This type of sampling is used to attempt to replicate the distributions of the underlying variable (annual consumption) in the comparison group.

This method is relatively simple and does a good job of matching the distributions of annual consumption in the middle of the consumption range (between 0.8 and 3 therms/day). However, as seen in Figures 5a and 5b, the consumption of the comparison group customers was lower than that of the treatment group in the lowest and highest bins. The source of this problem was identified as the large range of usage values in the smallest and largest deciles - up to 6 therms in the largest decile, compared to around 0.2 therms in the other deciles. Having such a wide range means that the match quality is significantly reduced. This problem could be mitigated by using finer percentile strata for sampling, although the number of available candidates within each bin might be reduced significantly. This bias results in relatively poor matches, as evidenced by Figures 5c and 5d.

Table 3. Decile bins used in stratified sampling



Bin endpoints	Range (Therms)
0	0.89
0.89	0.33
1.22	0.22
1.44	0.20
1.64	0.20
1.84	0.18
2.02	0.29
2.31	0.32
2.63	0.43
3.06	5.95
9.01	-

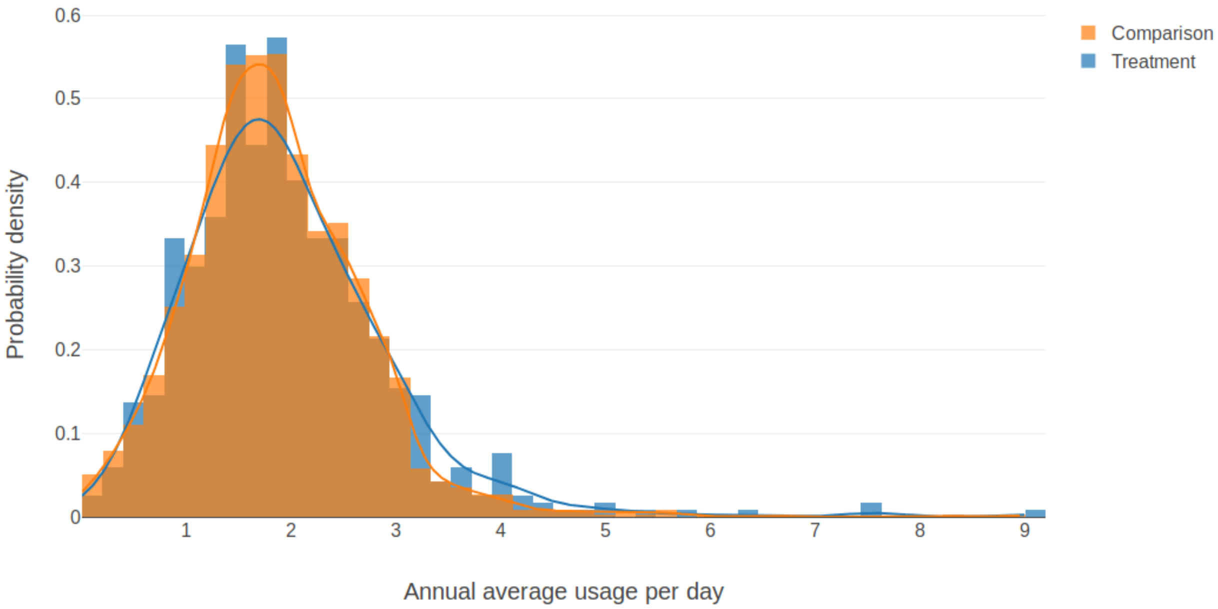


Figure 5a. Histograms of average annual gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined using stratified sampling ($n_{treatment}$ = 601; $n_{comparison}$ =2934).

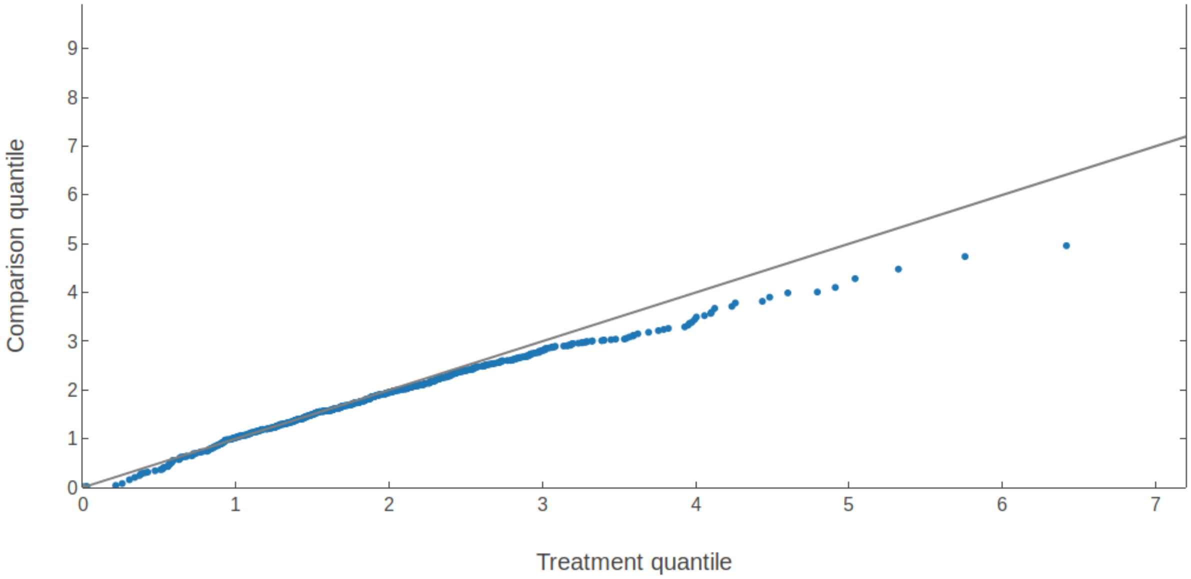


Figure 5b. Q-Q plot of annual average gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined using stratified sampling ($n_{treatment}$ = 601; $n_{comparison}$ =2934).

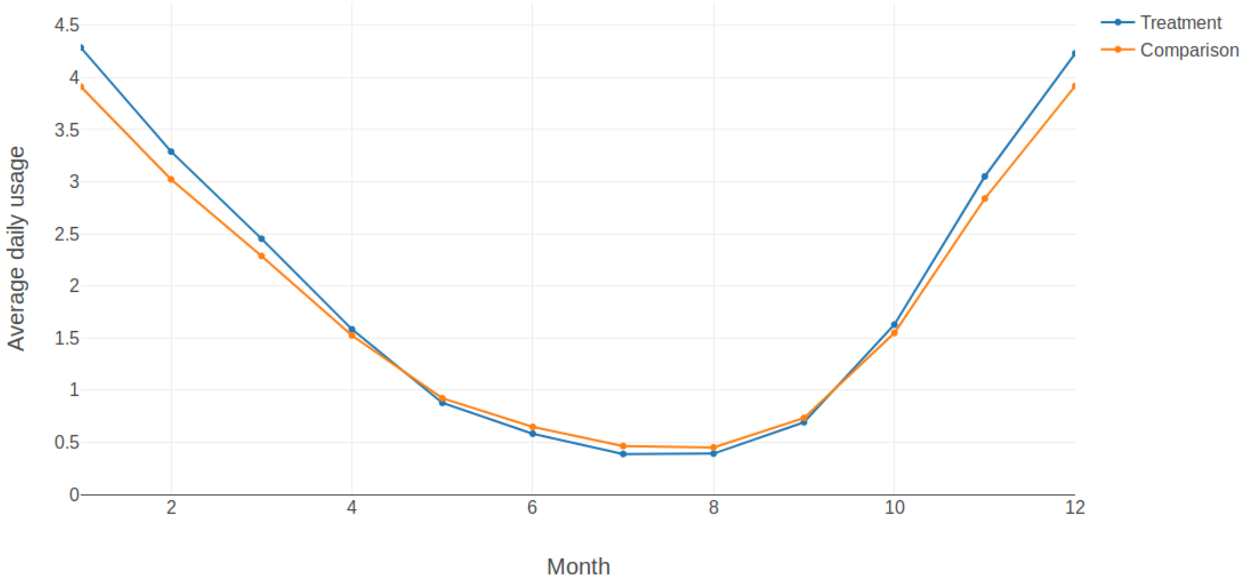


Figure 5c. Comparison of monthly average gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined using stratified sampling ($n_{treatment}$ = 601; $n_{comparison}$ =2934).

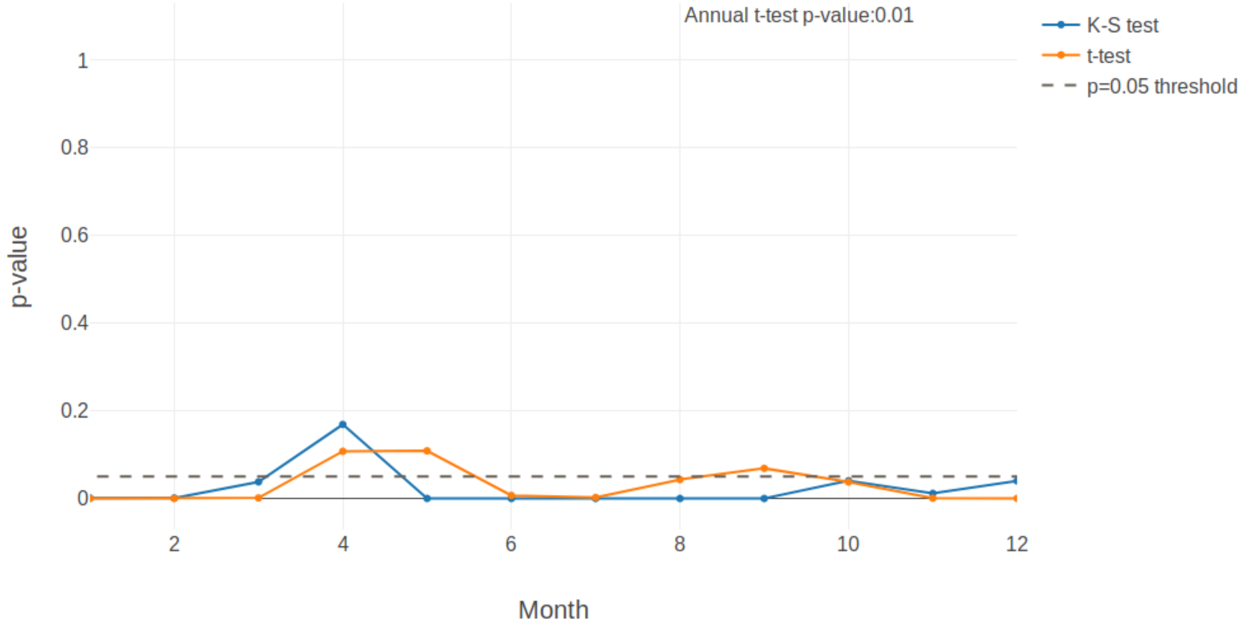


Figure 5d. Equivalence metrics for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined using stratified sampling. P-values larger than 0.05 indicate no significant differences in the mean (t-test) or distribution (K-S test) of monthly consumption ($n_{treatment}$ = 601; $n_{comparison}$ =2934).

2.6 Annual consumption matching

Another class of comparison group matching methods involves finding *n* matches for each treatment group member based on some distance metric calculated from one or more features in the baseline period. For example, comparison group matches can be selected by comparing the annual consumption and selecting *n* matches with the closest annual consumption. This requires slightly more computational resources than the previous portfolio-level methods, however, the matching step can be easily parallelized and scaled.

In the current test, we selected 5 nearest neighbors based on annual consumption alone. Figures 6a and 6b demonstrate that this method very precisely reproduced the distribution of annual consumption (which is intuitive, given that the matching is done on annual consumption). The algorithm was not able to identify close matches for a handful of treatment group customers with very high usage levels.

Unfortunately, the monthly consumption shown in Figure 6c reveals that the consumption of the comparison group exceeds that of the treatment group in summer and vice versa in winter. This would be concerning as the difference in baseline consumption is time-varying and implies that the seasonal variation in energy use is not equivalent in the two groups. The monthly K-S test shown in Figure 6d highlights this difference - while the match clearly passes the t-test at the annual level (p=0.91), the match quality is poor in January and May through September. Therefore, this method is not recommended for use, especially because the monthly matching method described in the following section provides much better match quality, with minimal additional complexity.

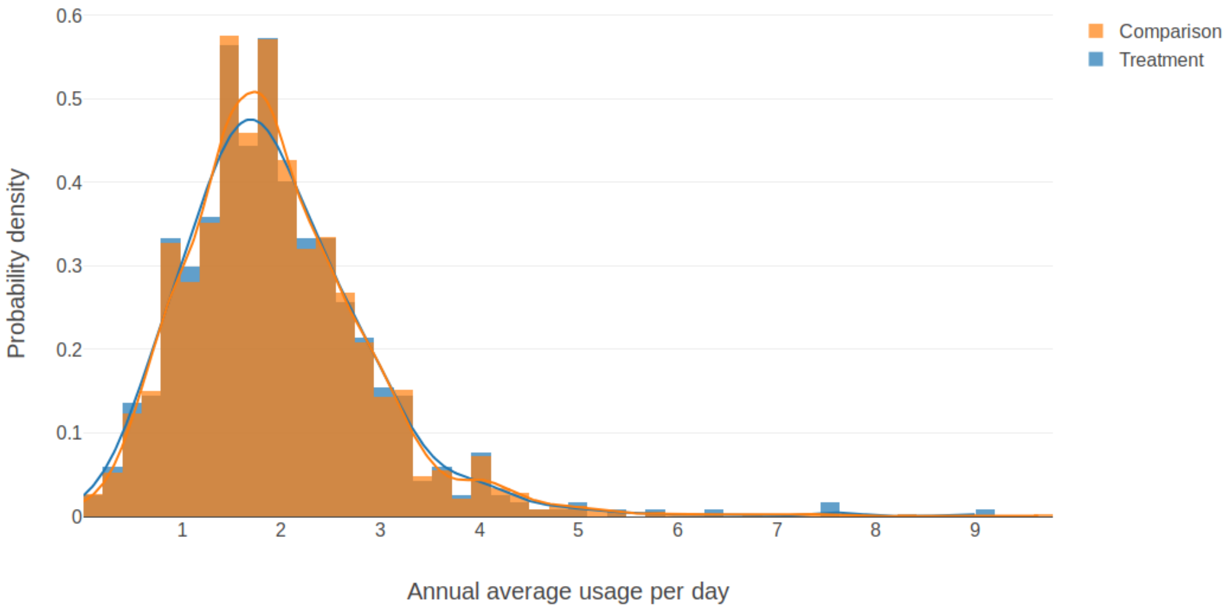


Figure 6a. Histograms of average annual gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined by individual matching on annual consumption ($n_{treatment}$ = 599; $n_{comparison}$ =2664).

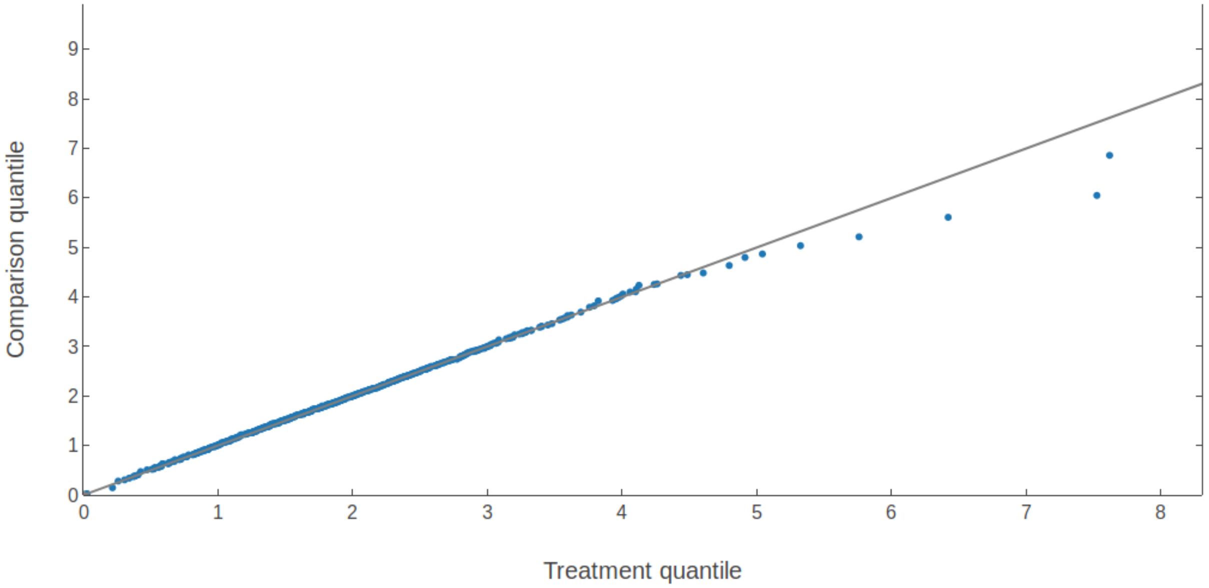


Figure 6b. Q-Q plot of annual average gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined by individual matching on annual consumption ($n_{treatment}$ = 599; $n_{comparison}$ =2664).

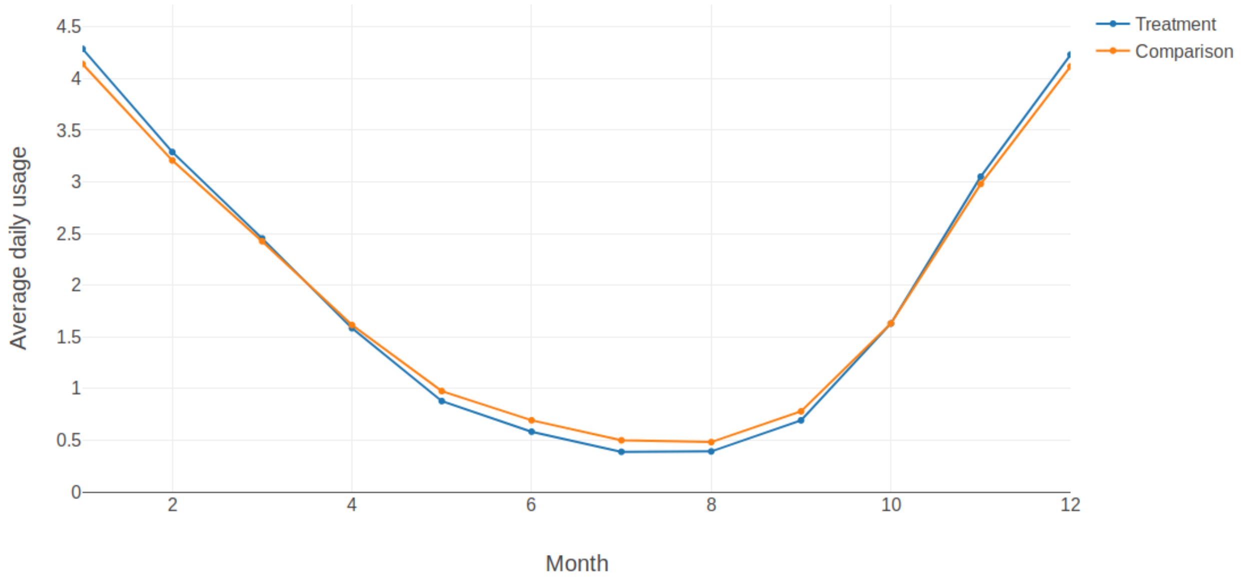


Figure 6c. Comparison of monthly average gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined by individual matching on annual consumption ($n_{treatment}$ = 599; $n_{comparison}$ =2664).

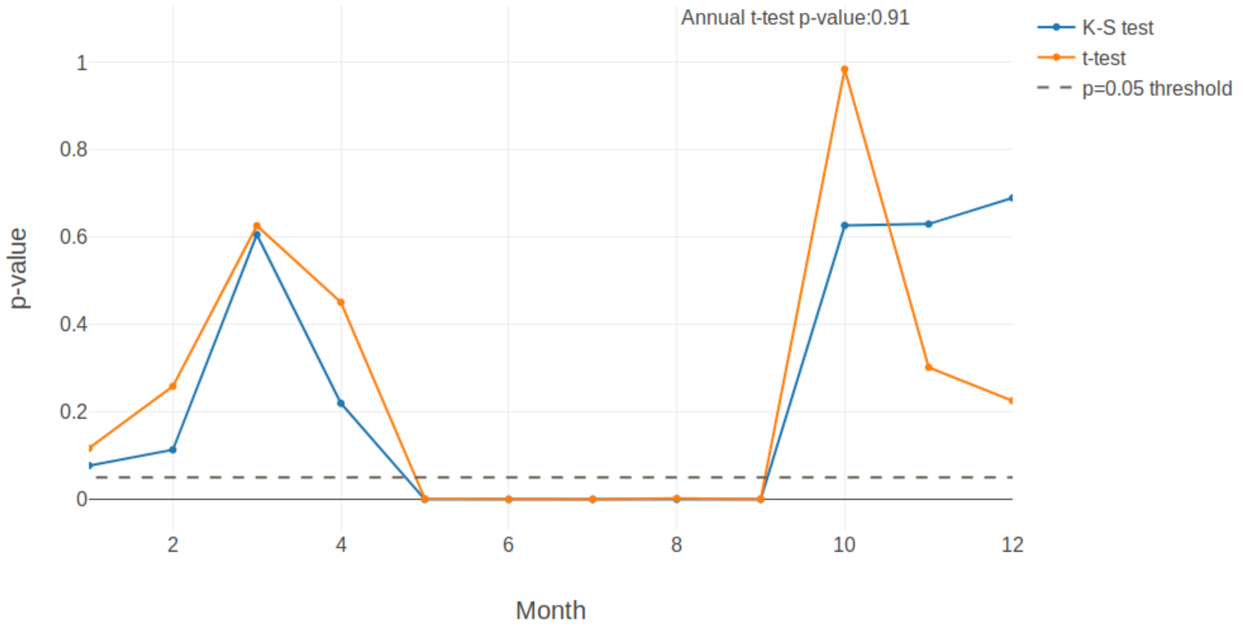


Figure 6d. Equivalence metrics for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined by individual matching on annual consumption. P-values larger than 0.05 indicate no significant differences in the mean (t-test) or distribution (K-S test) of monthly consumption ($n_{treatment}$ = 599; $n_{comparison}$ =2664).

2.7 Monthly consumption matching

In monthly consumption matching, comparison group is constructed by selecting *n* matches from the comparison group pool with the shortest distance *d* to the treatment group customer under consideration. The distance *d* is, in essence, a way to reduce 12 monthly consumption differences between any two customers to one metric (Figure 7ex). In the present study, we used Euclidean distance[6], for its simplicity and intuitiveness.

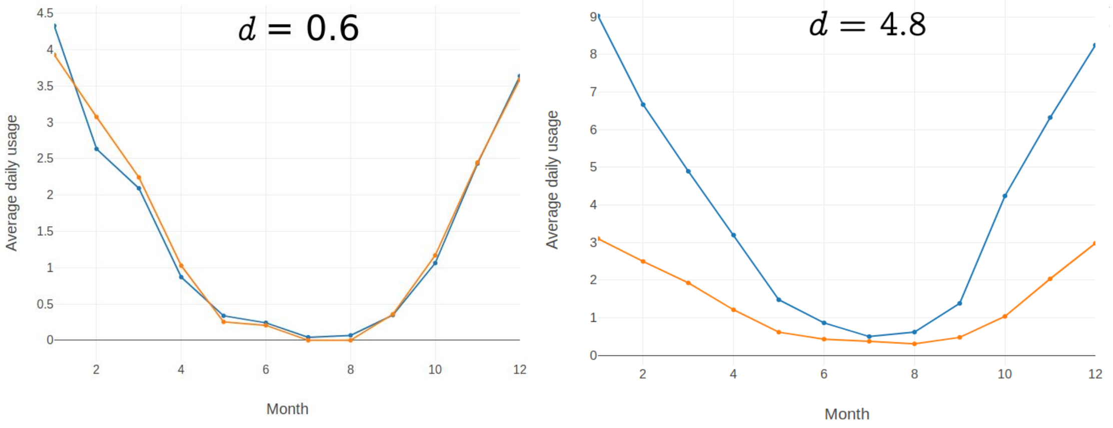


Figure 7ex. Reducing the monthly consumption traces for two customers to a single Euclidean distance metric *d*.

In the current test, we selected (without replacement) five nearest neighbors for each participant based on the Euclidean distance of monthly consumption. Figures 7a, b, c demonstrate that the comparison group very precisely replicates the distribution of annual and monthly consumption and all of the equivalence metrics indicate that this is a good match in the portfolio baseline year (Figure 7d).

The matching results for all methods are summarized in Section 4.

[6] https://en.wikipedia.org/wiki/Euclidean_distance

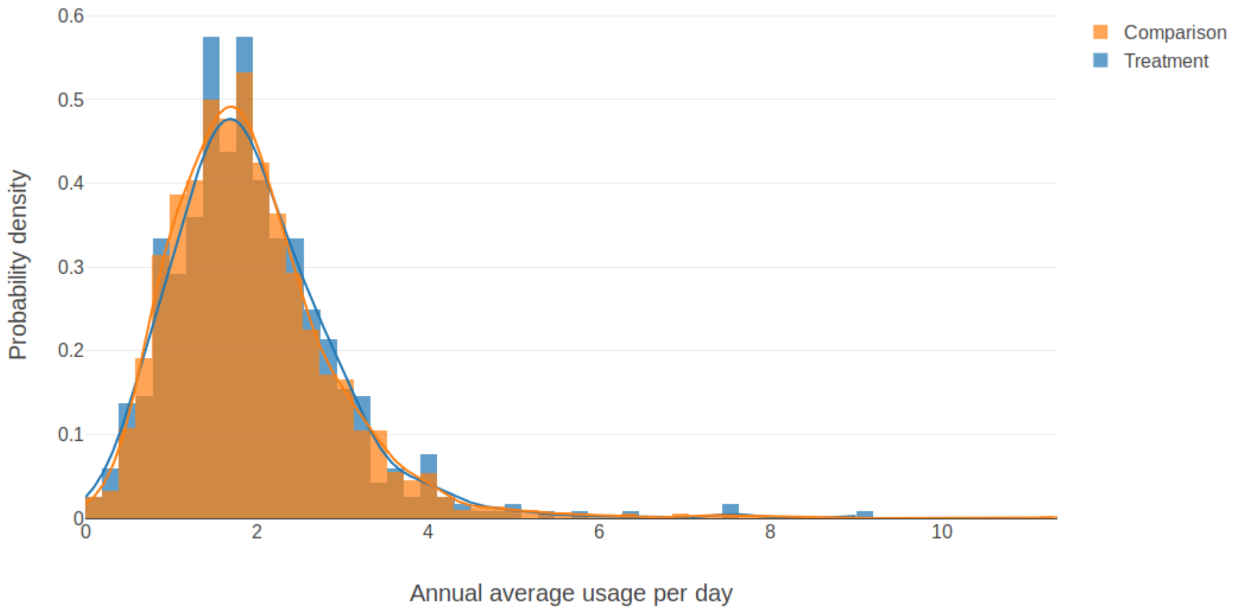


Figure 7a. Histograms of average annual gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ = 597; $n_{comparison}$ =2963).

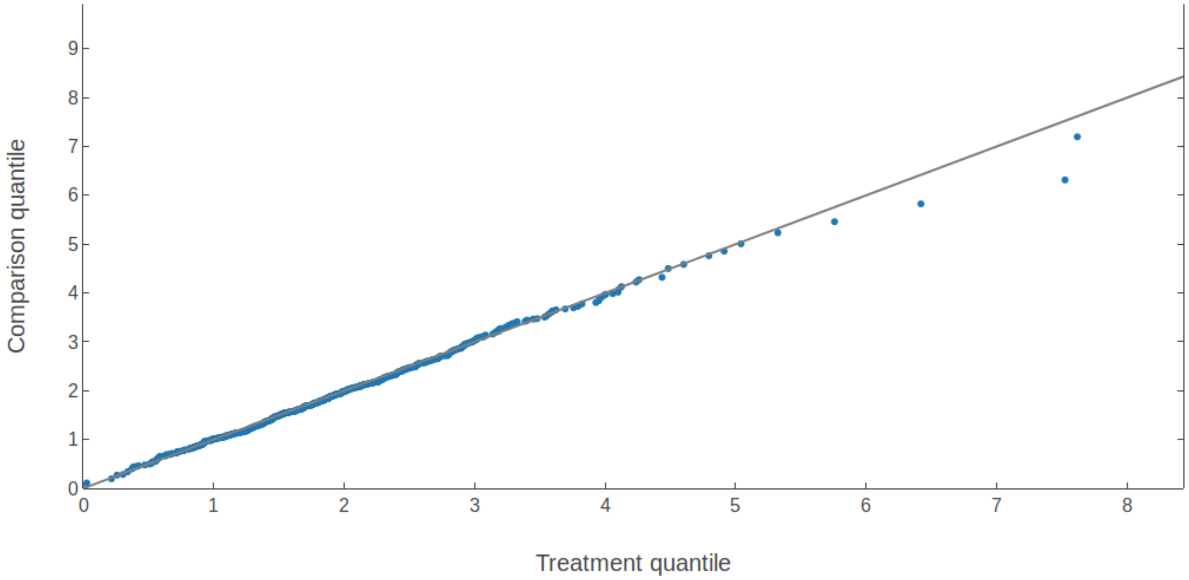


Figure 7b. Q-Q plot of annual average gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ = 597; $n_{comparison}$ =2963).

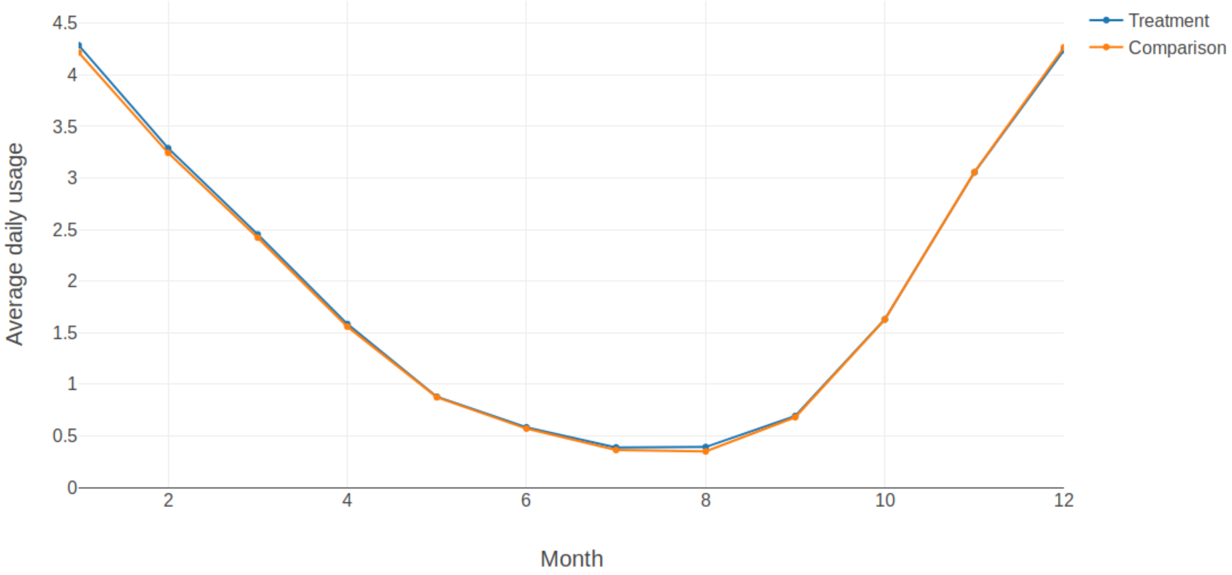


Figure 7c. Comparison of monthly average gas consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ = 597; $n_{comparison}$ =2963).

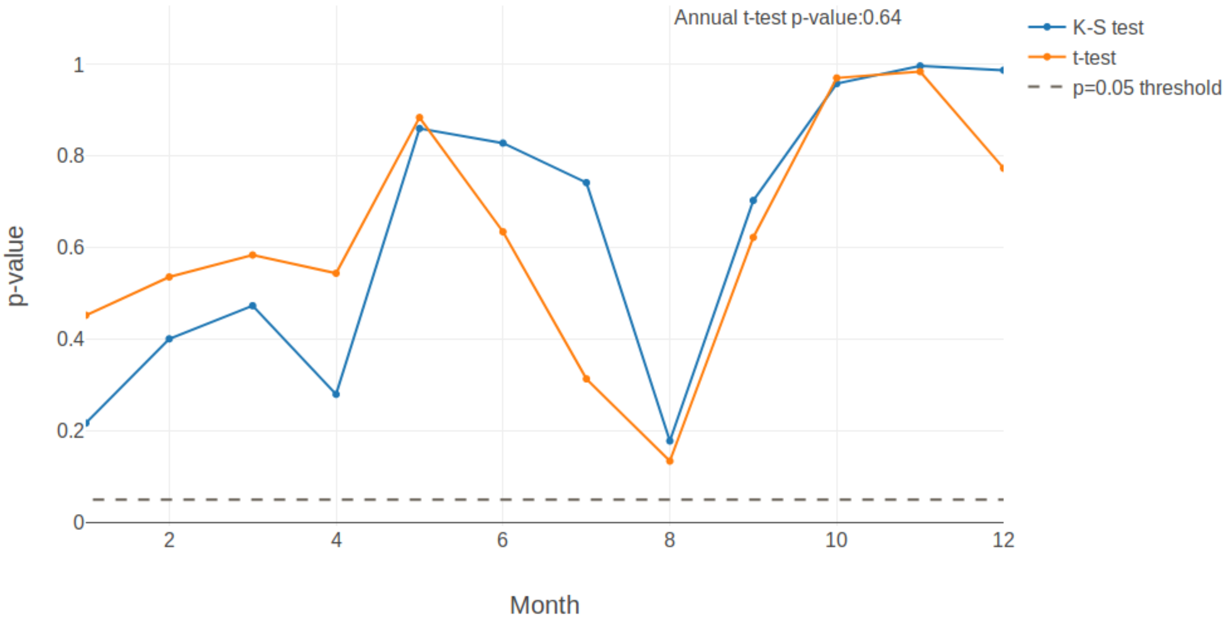


Figure 7d. Equivalence metrics for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined by individual matching on monthly consumption. P-values larger than 0.05 indicate no significant differences in the mean (t-test) or distribution (K-S test) of monthly consumption ($n_{treatment}$ = 597; $n_{comparison}$ =2963).

2.8 Out of sample results

The test dataset included consumption data for several years in the baseline period, especially for 2014 and 2015 program participants. This enabled an analysis of changes in energy consumption, absent any known energy efficiency interventions. For example, for 2014 program participants, the change in energy consumption between 2012 and 2013 should be close to that of the comparison group, if the two groups were well-matched (the difference between the two quantities should be theoretically equal to zero if the two samples are large enough and drawn from the same population).

Differences in Normalized Annual Consumption (DNAC) were calculated for several groups of projects and their corresponding comparison groups in baseline years immediately prior to the project year (for example, for 2015 participants, DNAC was calculated between 2013 and 2014). The difference in DNAC between the treatment and comparison groups (DDNAC), expressed as a percent of treatment group baseline NAC is shown in Figure 8:

$DDNAC_y = (NAC_{treatment,\ y-2} - NAC_{treatment,\ y-1}) - (NAC_{comparison,\ y-2} - NAC_{comparison,\ y-1})$

where y is the program participation year.

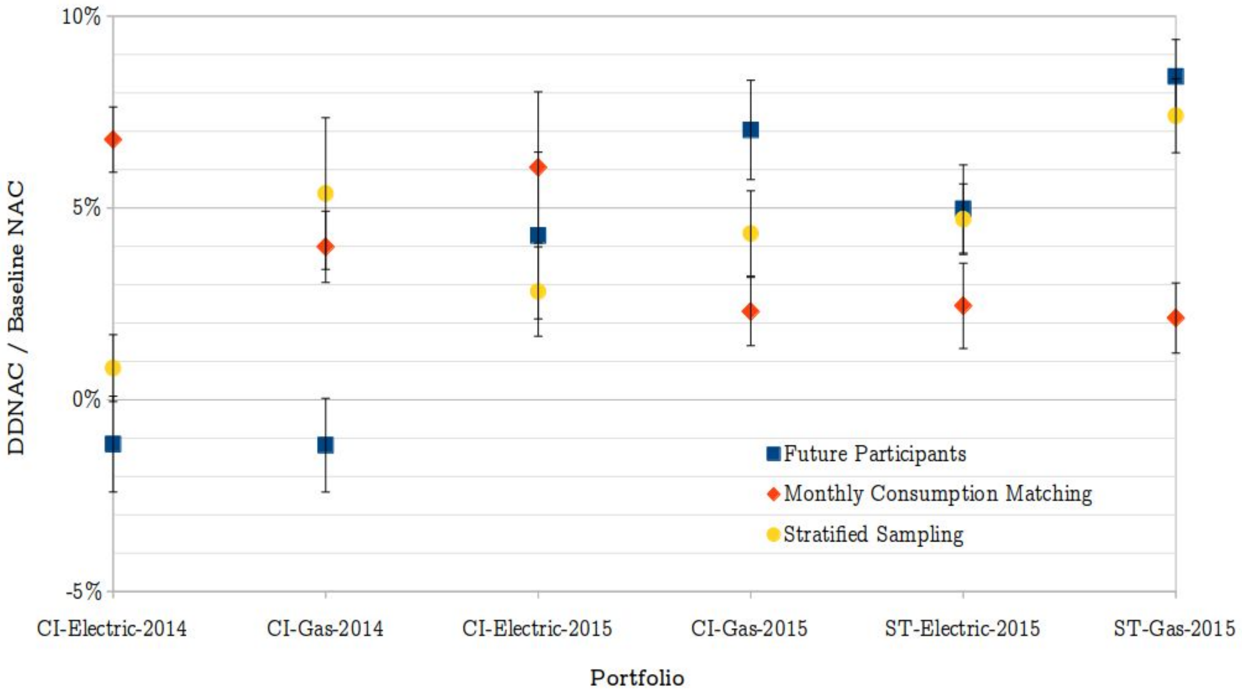


Figure 8. Baseline differences in Normalized Annual Consumption for different portfolios of projects (CI: Ceiling insulation, ST: Smart thermostats), using three different comparison group methods.

Table 4. Summary of out-of-sample testing datasets and results.

Measure	Program Year	Baseline year	Reporting year	Treatment group size*
Ceiling Insulation	2014	2012	2013	594 (Gas) 677 (Elec)
Ceiling Insulation	2015	2013	2014	604 (Gas) 719 (Elec.)
Smart thermostats	2015	2013	2014	432 (Gas) 425 (Elec.)

*Sample size attrition occurred for the out-of-sample tests due to insufficient baseline data.

				Sample size		DNAC		
Measure	**Year**	**Fuel**	**Comparison group**	**Treatment**	**Comparison**	**Treatment**	**Comparison**	**DDNAC/ Baseline NAC**
Ceil. Insulation	2014	Elec.	future	677	578	688	815	-1%
Ceil. Insulation	2014	Gas	future	594	624	27	35	-1%
Ceil. Insulation	2014	Elec.	matching	676	3075	686	-61	7%
Ceil. Insulation	2014	Gas	matching	594	2920	27	1	4%
Ceil. Insulation	2014	Elec.	stratified	677	2776	688	597	1%
Ceil. Insulation	2014	Gas	stratified	594	3006	27	-2	4%
Ceil. Insulation	2015	Elec.	future	719	611	735	256	4%
Ceil. Insulation	2015	Gas	future	604	478	19	-24	7%
Ceil. Insulation	2015	Elec.	matching	717	3473	733	56	6%
Ceil. Insulation	2015	Gas	matching	603	2973	19	5	2%
Ceil. Insulation	2015	Elec.	stratified	719	3374	735	133	5%
Ceil. Insulation	2015	Gas	stratified	604	3157	19	-10	5%
Thermostat	2015	Elec.	future	425	1474	544	77	5%
Thermostat	2015	Gas	future	432	1558	35	-21	8%
Thermostat	2015	Elec.	matching	425	2095	544	314	2%
Thermostat	2015	Gas	matching	432	2151	35	21	2%
Thermostat	2015	Elec.	stratified	425	1965	544	279	3%
Thermostat	2015	Gas	stratified	432	2139	35	-14	7%

One overarching observation is that, in almost cases, the treatment group saves more energy in the year immediately prior to the intervention than the corresponding comparison group over that same period. While this would require further investigation, it appears to support the hypothesis that program participants have a higher likelihood of undertaking energy efficiency upgrades in the year leading to program participation. This would imply that such an analysis may be more informative if done further back in time, which was not possible for data availability reasons in this case.

Another interesting observation is that none of the methods provide consistently good results for all portfolios of projects (measure/year combinations). The DNAC of the comparison groups from all three methods are within 5% of the treatment group in most cases, however, in some cases, the discrepancies are much worse for particular methods with particular portfolios. While these results are inconclusive, they suggest that the baseline period DDNAC may serve as an additional quality metric for the comparison groups. Calculating this quantity using different comparison group methods could allow matching problems to be discovered early on. However, the numerical values of baseline DDNAC should be interpreted carefully, due to the apparent bias of the treatment DNAC in the year prior to program participation.

3. Other methodological issues

Other issues related to data handling and savings calculations that were investigated in this study are presented in this section.

3.1 Baseline/reporting period length

The length of the baseline and reporting periods that are included in the savings models may affect results in two ways:

- Periods that are too short may not capture the full range of independent variables (weather) that are typically experienced.
- Periods that are too long increase the chances of unexpected changes in a building's energy use (e.g. due to a change in occupancy or in the building's equipment).

It is generally agreed that a minimum of 12 months of billing data should be used in order to capture at least one annual cycle of energy use. However, there are no general guidelines about the maximum length of time to include in savings analysis. As part of the CalTRACK 2.0 updates,[7] site-level regression models were fit to 1000 program participants from a number of different programs in Oregon, only varying the length of the baseline period that was used to fit the models- between 12 and 24 months in 3-month increments. As shown in Figure 8, no monotonic trends were obvious in the normalized annual consumption, however, there were some cyclical trends, likely corresponding to the model being weighted towards the seasons with more data. While these are small variations in NAC, they may translate to large biases in savings, especially for measures with small savings relative to the baseline.

Since the predicted baseline may be unstable with different baseline period lengths, which may, in turn, affect calculated savings, the consensus of the CalTRACK 2.0 working group was to set the maximum baseline period at 12 months, since the year leading to the energy efficiency intervention is the most indicative of recent energy use trends and prolonging the baseline period increases the chance of other unmeasured factors affecting the baseline.

[7] https://github.com/CalTRACK-2/caltrack/issues/68

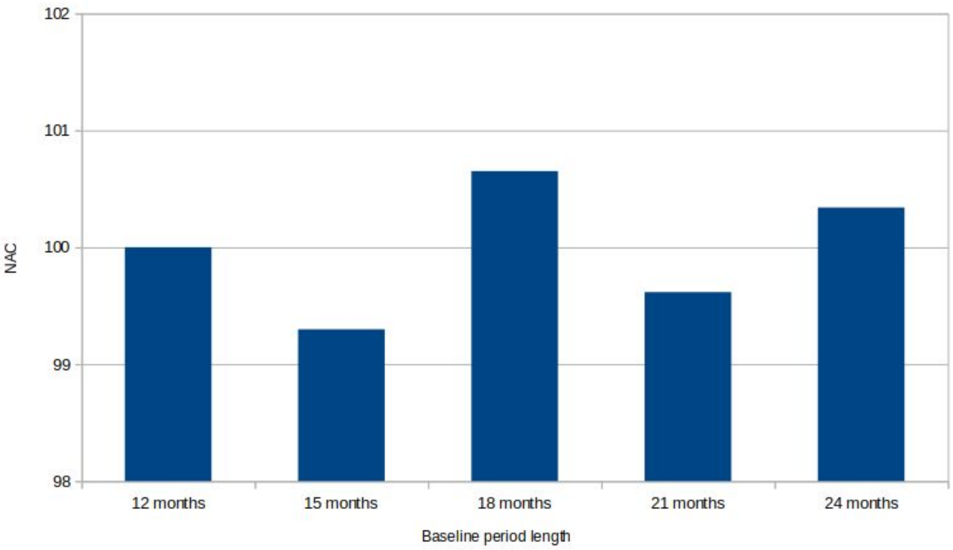


Figure 9. Effect of baseline period length on normalized annual consumption using billing data. Y axis (Baseline Normalized Annual Consumption) is in percent.

3.2 Project blackout period

The blackout period refers to the time period between the end of the baseline period and the beginning of the reporting period, which is used to calculate savings using site-level regression. Typically, this is specified to coincide with the project installation time period, however, it may be prolonged if needed. For example, the blackout period could be set as the entire program year, with the baseline period set as the calendar year preceding the program year and the reporting period set as the calendar year following the program year. A loose blackout period (entire program year) is more convenient from a data preparation perspective and requires less data collection (project dates are not required), while a tight blackout period (spanning project installation dates) ensures that the most recent usage data is used in modeling and would ensure more data availability (only 2 years of data required for most projects vs 3 years for the loose blackout). As an example, for a project that was installed between 4/5/2014 and 4/10/2014, these periods are as follows:

- Tight blackout:
    - Baseline: 4/5/2013 - 4/4/2014
    - Reporting: 4/11/2014 - 4/10/2015
- Loose blackout:
    - Baseline: 1/1/2013 - 12/31/2013
    - Reporting: 1/1/2015 - 12/31/2015

For each comparison site, the baseline and reporting periods were set to be the same as those for the corresponding matched treatment customer.

As part of the current tests, both of these approaches were tested with the monthly consumption matching method and the savings results are presented in Table 4. In general, there was no significant difference in data processing or computational requirements. However, there is an obvious negative bias in energy use change for the treatment and comparison groups when using a loose blackout period, while the overall savings turned out to be larger. In general, it is expected that more recent consumption values are more indicative of recent energy use trends and can build a more representative baseline model. Therefore, we are recommending the use of a tight blackout period, defined by the project installation dates, unless the exact project installation dates are unknown, in which case a loose blackout may be used. In either case, the same methodology should be applied consistently to all customers in both the treatment and comparison groups.

3.3 Geographical screen

The first step when matching comparison group members with program participants is to construct a comparison group candidate pool based on similar geography, customer type, building type etc. For the residential use case, the main filter to apply is the geographical screen where the pool of comparison group candidates is limited to those in the same geographical area as the corresponding program participant to which they are being matched. The geographical screen can be applied at different levels of spatial granularity (e.g. zip code vs weather station vs county vs climate zone vs state).

The purpose of this task was to determine the level of granularity that is most suitable for comparison group matching. The group equivalence was compared using the monthly consumption matching method at the state, county and zip code levels. Figures 10a-d demonstrate that even with the most granular geographical screen, this method was still able to find good matched groups, albeit the quality of the match reduces slightly with higher granularity. The main drawback when using more granular geography is the reduction of the size of the comparison group candidate pool, which may mean that the optimal match based on consumption is not necessarily identified. On the other hand, consumption is not the only factor affecting energy use patterns, and using granular geographical screens may help equalize other factors, for example, building stock and demographic characteristics. We are, therefore, recommending to use a zip code-level or, alternatively, a weather station screen when applying the monthly consumption matching method.

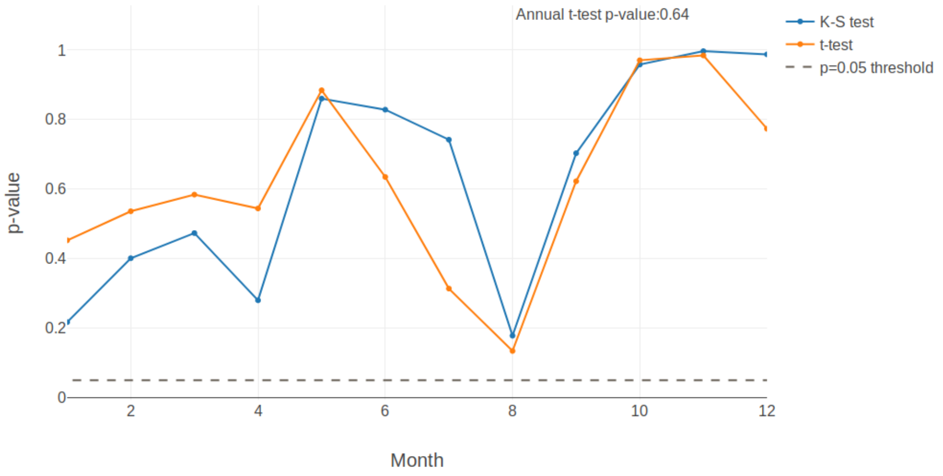


Figure 10a. Equivalence metrics for the treatment group (Ceiling insulation, 2014 program year) and a comparison group from the same state determined by individual matching on monthly consumption.

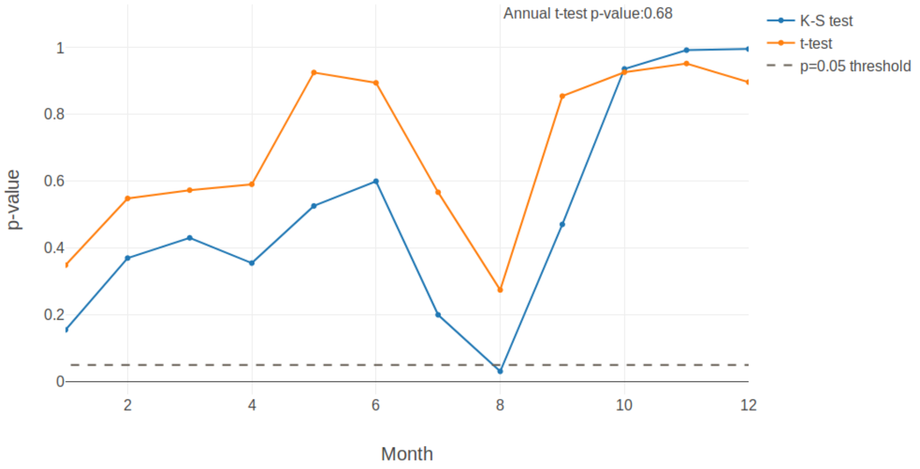


Figure 10b. Equivalence metrics for the treatment group (Ceiling insulation, 2014 program year) and a comparison group from the same county determined by individual matching on monthly consumption.

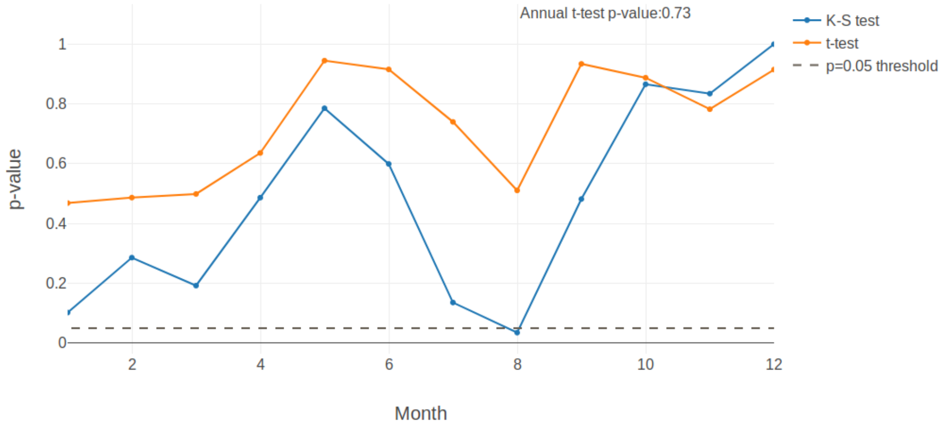


Figure 10c. Equivalence metrics for the treatment group (Ceiling insulation, 2014 program year) and a comparison group mapped to the same weather station determined by individual matching on monthly consumption.

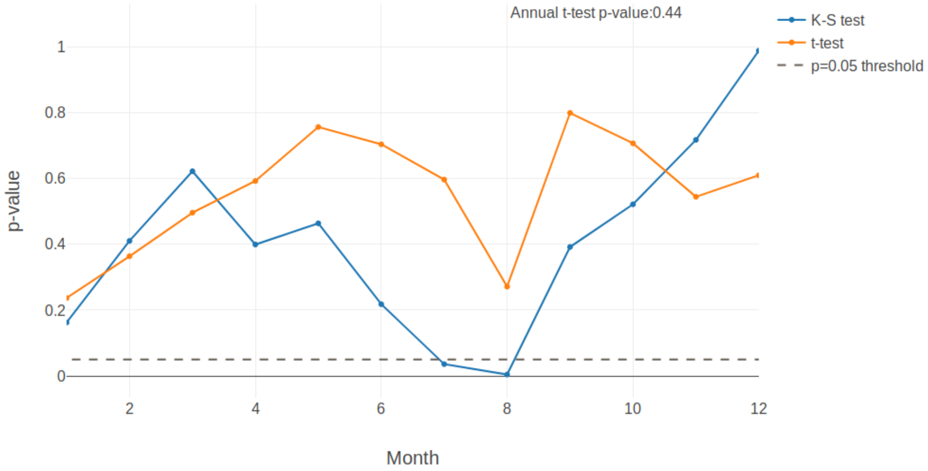


Figure 10d. Equivalence metrics for the treatment group (Ceiling insulation, 2014 program year) and a comparison group from the same zip code determined by individual matching on monthly consumption.

Side note on matching with gas data

One interesting observation in Figure 10d, is that the monthly consumption matching method at the zip code level fails the K-S test for the months of July and August. So, we drilled in to the data to determine if this was cause for concern. What we discovered was that, because the vast majority of residential customers use little or no gas in the summer months, the distribution of their monthly consumption is highly skewed in those months (Figure 11a). The K-S test captures differences in cumulative distributions, which means that even small differences in the number of customers using little or no gas between the treatment and comparison groups (the leftmost bar in Figure 11a, right) can translate to significant difference in the cumulative distributions (Figure 11b, right), causing the match to fail the monthly K-S test. Overall, this indicates that it should be acceptable for a match on gas data to fail the K-S test for 2 or 3 summer months, as long as the corresponding t-test results are acceptable. The same issue does not appear to occur with electricity data.

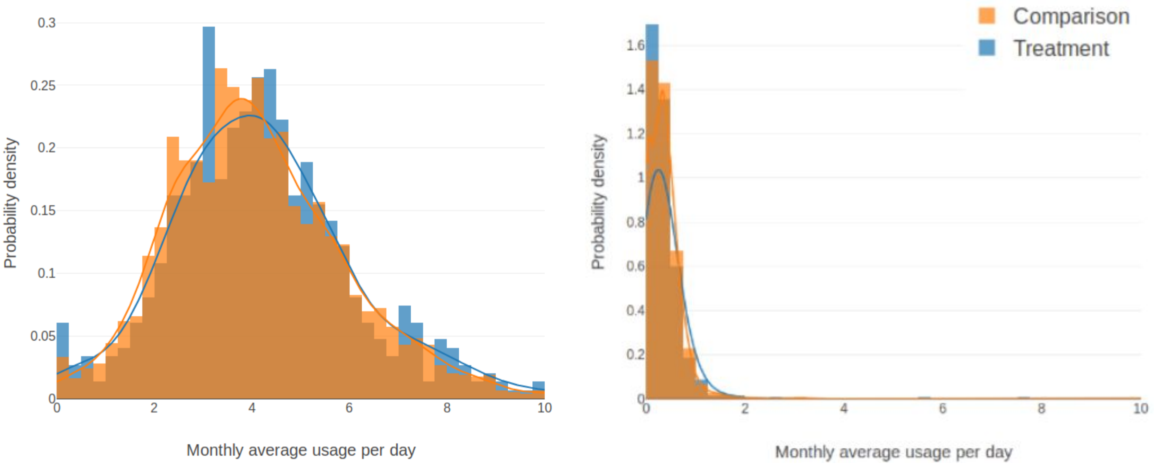


Figure 11a. Histograms of monthly average gas consumption in January 2013 (left) and July 2013 (right).

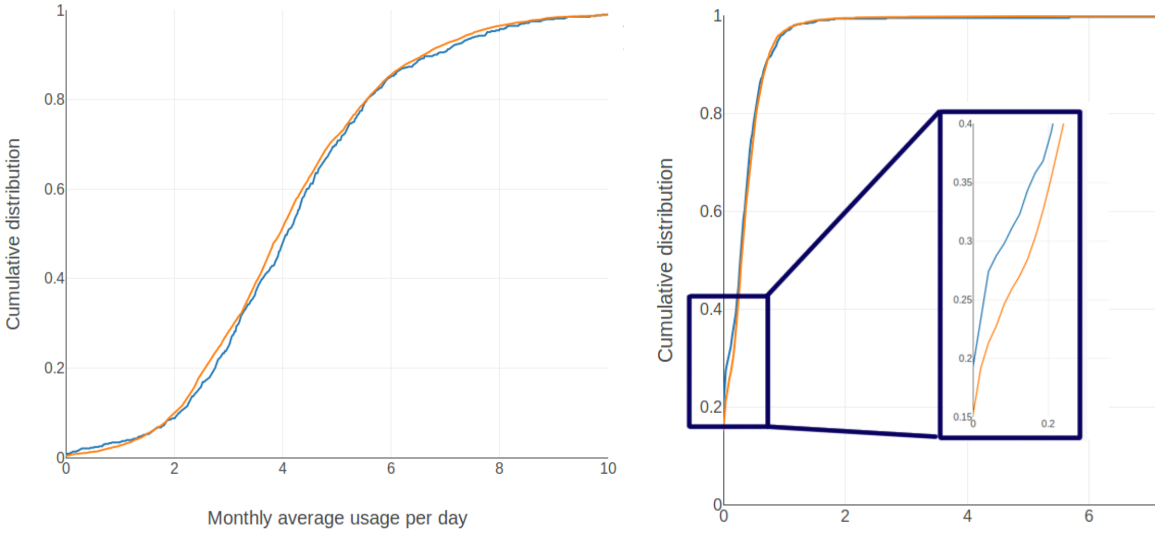


Figure 11b. Cumulative distributions of monthly average gas consumption in January 2013 (left) and July 2013 (right).

3.4 Sampling with/without replacement

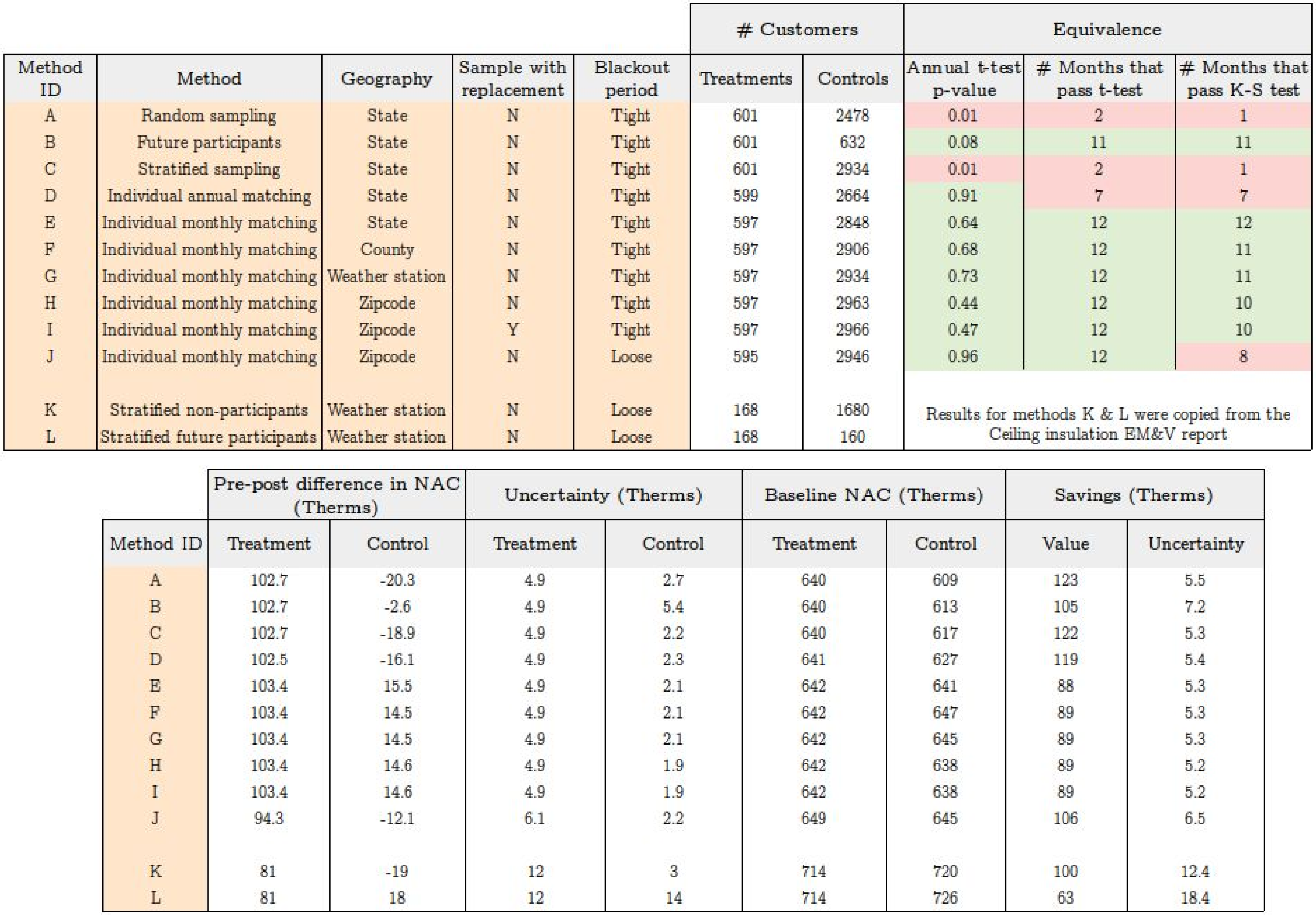
Sampling and matching with replacement is usually required when the available pool of comparison group candidates is limited. While this is not the case with Energy Trust data, we tested an additional scenario using the monthly consumption matching method to verify any potential issues. As expected, sampling with replacement only changed approximately 5% of matches in the comparison group and had an almost imperceptible effect on savings results. This issue may require further investigation for much smaller datasets, but for now, we are recommending proceeding with sampling without replacement as a default.

4. Savings Analysis

A summary of matching results is shown in Table 5. Overall, it appears that the future participant group and individual customer matching based on monthly consumption consistently pass all of the equivalence tests we have selected for this study. It must be noted that comparison group matching is dataset-specific and must, therefore, be repeated for new measures and analysis time periods. Treatment group sample sizes vary slightly because if a treatment customer has no matches with valid savings estimates (usually due to insufficient baseline data), the treatment is also dropped from the analysis.

Savings results were also computed for all cases using site-level CalTRACK 2.0 methods (Table 5 and Figure 12). Results from a recent EM&V study conducted by Energy Trust are included in the table. The results are not directly comparable, since the customer data in the Energy Trust study underwent more intensive data cleaning, the sample sizes are much smaller and certain methodological choices are different, however methods K and L are most comparable to methods C and B, respectively. The usage ranges of the groups are also different as evidenced by the baseline NAC. Nevertheless, the results of the monthly consumption matching method appears to be reasonably close. For example, the difference in NAC is very similar to that estimated for the stratified future participant group used by Energy Trust (15 +/-2 vs. 18 +/- 14 therms). The savings net of exogenous trends from the monthly consumption method amounted to 89 +/- 5 therms, while Energy Trust reported 81 +/- 22 therms based on the midpoint between a future participant and a non-participant group. The difference in NAC for the treatment group is somewhat different (103 +/-5 therms vs. 81 +/-12 therms), but these differences can be attributed to the different group compositions as well as the use of a loose blackout period in the Energy Trust study. For example, when using a loose blackout with the monthly consumption matching method, the difference in NAC for the treatment group drops to 94 +/-6 therms, much closer to the Energy Trust estimate.

Table 5. Summary of treatment/comparison group equivalence (top) and normalized savings results (bottom)



Method ID	Method	Geography	Sample with replacement	Blackout period	# Customers		Equivalence		
					Treatments	Controls	Annual t-test p-value	# Months that pass t-test	# Months that pass K-S test
A	Random sampling	State	N	Tight	601	2478	0.01	2	1
B	Future participants	State	N	Tight	601	632	0.08	11	11
C	Stratified sampling	State	N	Tight	601	2934	0.01	2	1
D	Individual annual matching	State	N	Tight	599	2664	0.91	7	7
E	Individual monthly matching	State	N	Tight	597	2848	0.64	12	12
F	Individual monthly matching	County	N	Tight	597	2906	0.68	12	11
G	Individual monthly matching	Weather station	N	Tight	597	2934	0.73	12	11
H	Individual monthly matching	Zipcode	N	Tight	597	2963	0.44	12	10
I	Individual monthly matching	Zipcode	Y	Tight	597	2966	0.47	12	10
J	Individual monthly matching	Zipcode	N	Loose	595	2946	0.96	12	8
K	Stratified non-participants	Weather station	N	Loose	168	1680	Results for methods K & L were copied from the Ceiling insulation EM&V report		
L	Stratified future participants	Weather station	N	Loose	168	160			

	Pre-post difference in NAC (Therms)		Uncertainty (Therms)		Baseline NAC (Therms)		Savings (Therms)	
Method ID	Treatment	Control	Treatment	Control	Treatment	Control	Value	Uncertainty
A	102.7	-20.3	4.9	2.7	640	609	123	5.5
B	102.7	-2.6	4.9	5.4	640	613	105	7.2
C	102.7	-18.9	4.9	2.2	640	617	122	5.3
D	102.5	-16.1	4.9	2.3	641	627	119	5.4
E	103.4	15.5	4.9	2.1	642	641	88	5.3
F	103.4	14.5	4.9	2.1	642	647	89	5.3
G	103.4	14.5	4.9	2.1	642	645	89	5.3
H	103.4	14.6	4.9	1.9	642	638	89	5.2
I	103.4	14.6	4.9	1.9	642	638	89	5.2
J	94.3	-12.1	6.1	2.2	649	645	106	6.5
K	81	-19	12	3	714	720	100	12.4
L	81	18	12	14	714	726	63	18.4

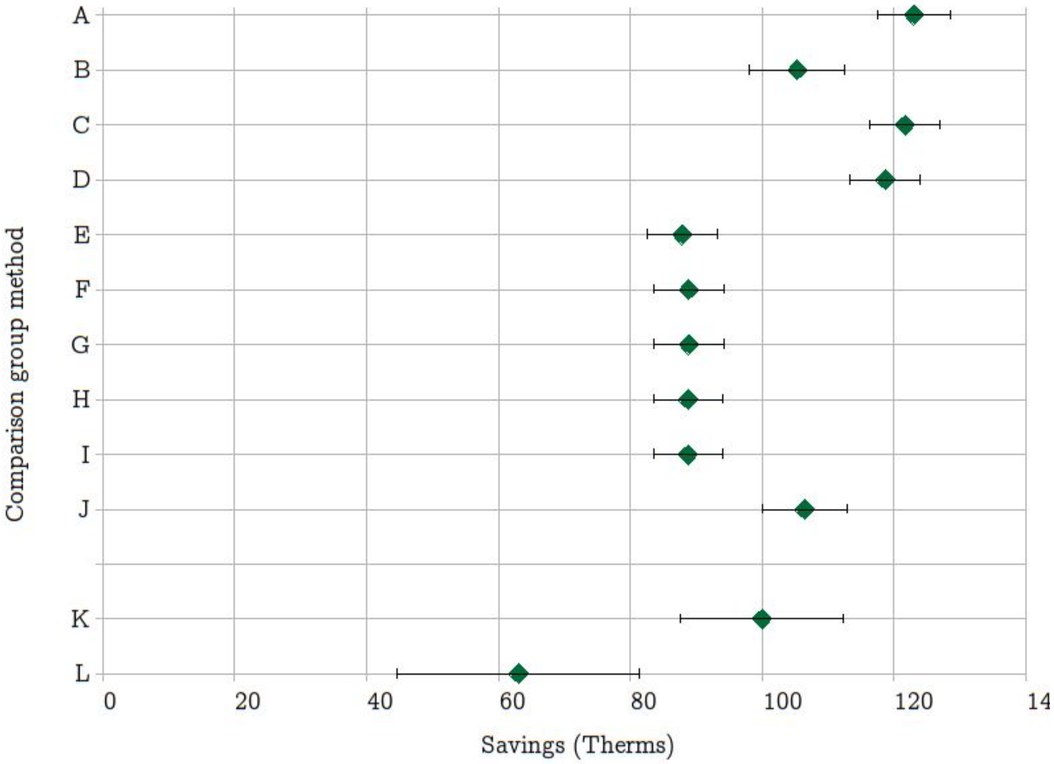


Figure 12. Comparison of savings results for different comparison group matching methods.

5. Recommendations

The purpose of this study was to develop a set of automated procedures for impact evaluation. These procedures are meant to reduce the resource and time requirements for impact evaluation and to enable more timely program feedback. This study tested several standard methods of comparison group identification, along with several methodological choices that are used in implementation. It is clear from Section 2 that different methods can yield different results using the same datasets. And it is unclear if there is one "best" method - in particular, monthly consumption matching at the zipcode level and future participant groups offer similar levels of performance. Therefore, our primary recommendation when implementing automated comparison group identification is to **automate the calculation, not the interpretation of results**. Moreover, this first version of comparison group methods will be continuously improved as they are used with more varied datasets and as we gain more experience with their strengths and limitations. More specific recommendations for different use cases are outlined below.

General recommendations for comparison group identification

The following methodological approaches were found to yield consistent matching results:

- Use several metrics to judge the quality of a match: e.g. t-test on annual consumption, t-test and K-S test on monthly consumption and the difference in differences of normalized annual consumption in two baseline years (if enough data is available).

- Construct the comparison group candidate pool from customers who are known not to have participated in any energy efficiency programs during the analysis period. Select comparison group candidates located in the same zip code as the corresponding participant.

- When a customer from the comparison group candidate pool is matched to a participant, discard it from the pool (sampling without replacement).

Standard program impact evaluation use case

The first phase of an impact evaluation is to determine changes in energy consumption for program participants, controlling for atypical weather and exogenous trends in energy consumption. These evaluations are typically performed 12 months or more following the end of a program.

- Use up to three methods of comparison group identification depending on data availability: individual customer matching on monthly consumption with Euclidean distance as the matching metric, stratified sampling of future participant groups and stratified sampling of past participant groups. One feasible option for calculating results, that may require further investigation, is to ensemble the results of multiple methods (e.g. use the average DNAC for several comparison groups).

- Screen out customers that do not meet CalTRACK data sufficiency guidelines, as well as outliers in terms of consumption data. As initial guidance, we recommend removing the top and bottom 0.5% of treatment group customers by annual consumption.

- For savings calculations, limit the baseline and reporting periods to 12 months and use a tight blackout period, if the project dates are known.

Pay-for-performance impact evaluation use case

Settlement in pay-for-performance programs requires complete transparency in the methods used to calculate payable savings. Payment can be based solely on the difference in energy consumption, normalized for weather and/or other routine variables (e.g. occupancy). In that case, comparison group identification is required for evaluation purposes only and the recommendations for standard program impact evaluation may be followed.

On the other hand, if comparison groups are to be factored into payments, we recommend that the comparison group identification method be contractually set before the launch of a pay-for-performance procurement and accommodated in the program design. For example, the procurement contract may specify a single method (e.g. monthly consumption matching), or a set procedure (e.g. ensemble average of three methods). Several comparison group identification methods only require baseline data, and can be used even before projects are completed at participating sites. These may be preferred in pay-for-performance settings to allow aggregators to track the performance of the comparison group alongside that of program participants. Portfolio-level comparison group identification (e.g. stratified sampling, past participant groups) may be harder to implement on an ongoing basis as it requires participants to be registered in cohorts that are individually tracked. In general, it is essential for both parties (procurer and aggregator) to agree to a practical method for integrating comparison group in the payment model, and for the relevant data to be made available as well.

Appendix

This appendix includes matching results for datasets other than the primary one discussed in the body of the report. Details of these datasets are included in Table 1 as well as the figure captions.

A1. Equivalence metrics for Ceiling insulation in 2013 (Gas)

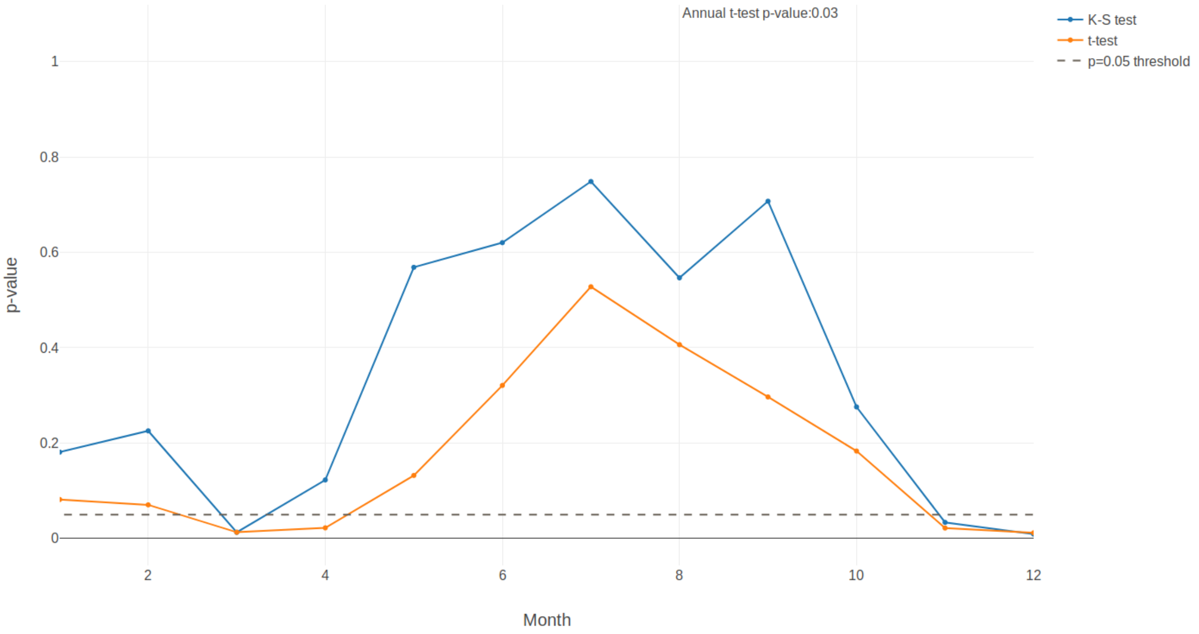


Figure A1a. Equivalence metrics for the treatment group (Ceiling insulation, 2013 program year) and a future participant comparison group (Ceiling insulation, 2014 program year) ($n_{treatment}$ =637; $n_{comparison}$ =591).

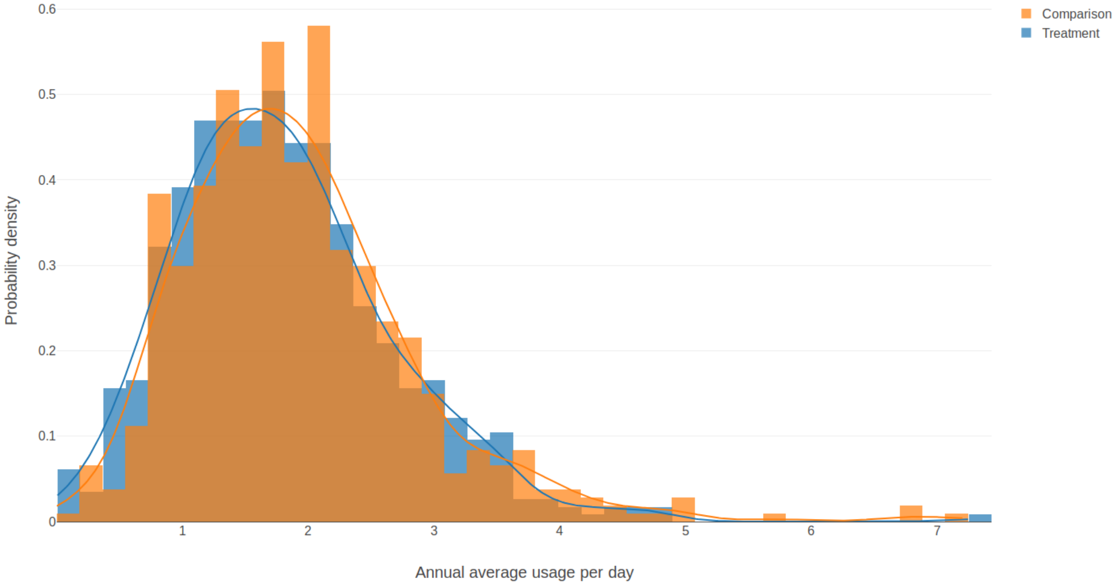


Figure A1b. Histograms of average annual gas consumption per day for the treatment group (Ceiling insulation, 2013 program year) and a future participant comparison group ($n_{treatment}$ =637; $n_{comparison}$ =591).

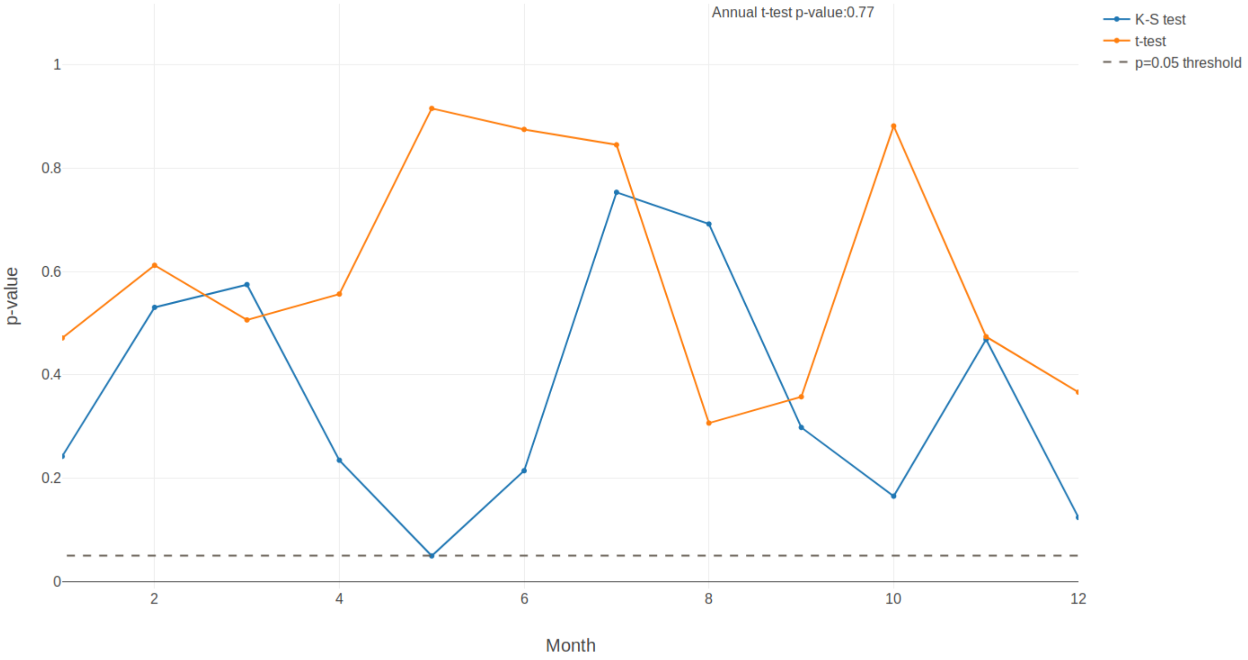


Figure A1c. Equivalence metrics for the treatment group (Ceiling insulation, 2013 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =631; $n_{comparison}$ =3138).

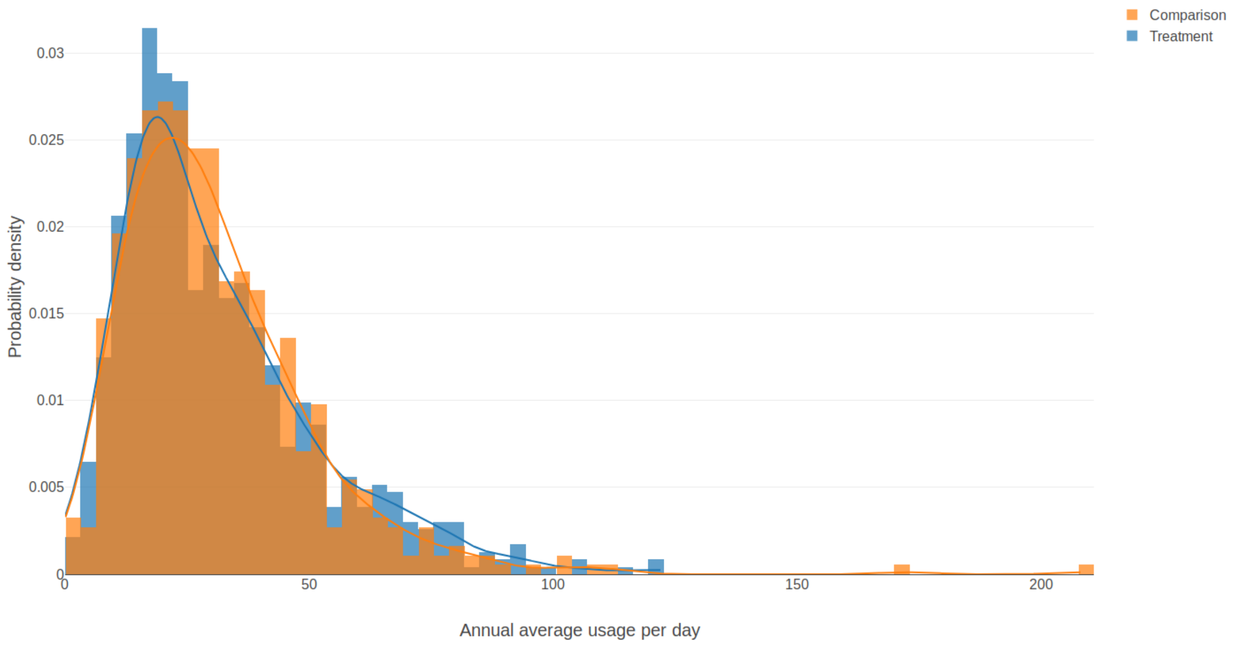


Figure A1d. Histograms of average annual gas consumption per day for the treatment group (Ceiling insulation, 2013 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =631; $n_{comparison}$ =3138).

A2. Equivalence metrics for Ceiling insulation in 2013 (Electricity)

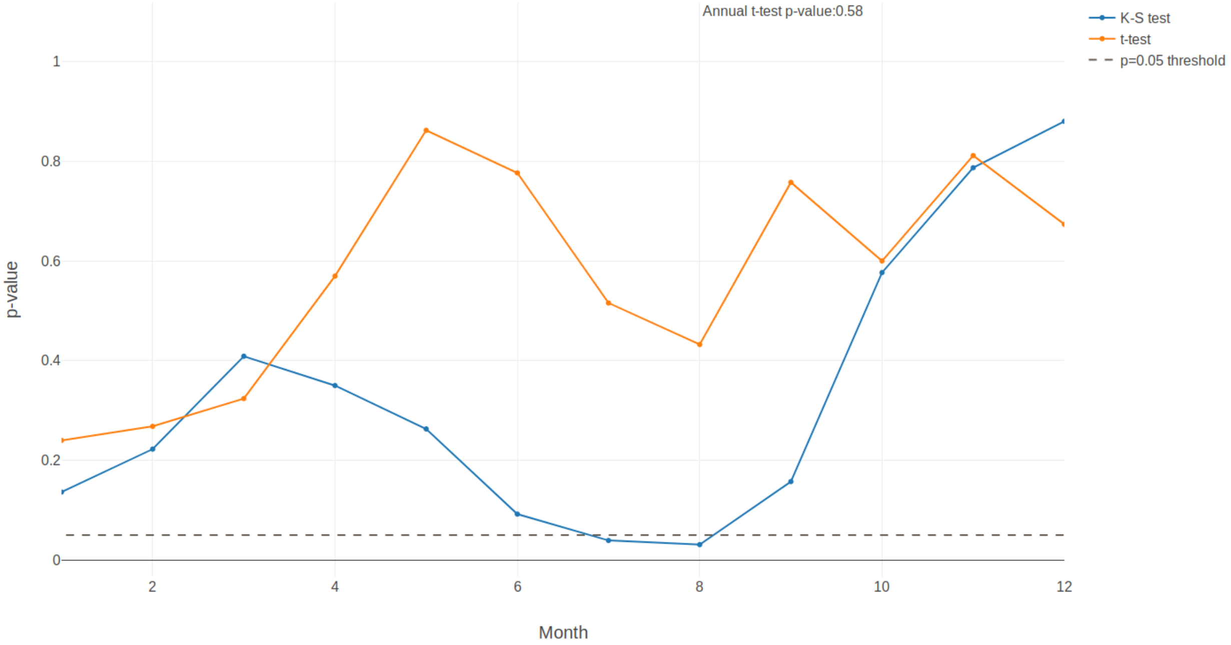


Figure A2a. Equivalence metrics for the treatment group (Ceiling insulation, 2013 program year) and a future participant comparison group (Ceiling insulation, 2014 program year) ($n_{treatment}$ =750; $n_{comparison}$ =584).

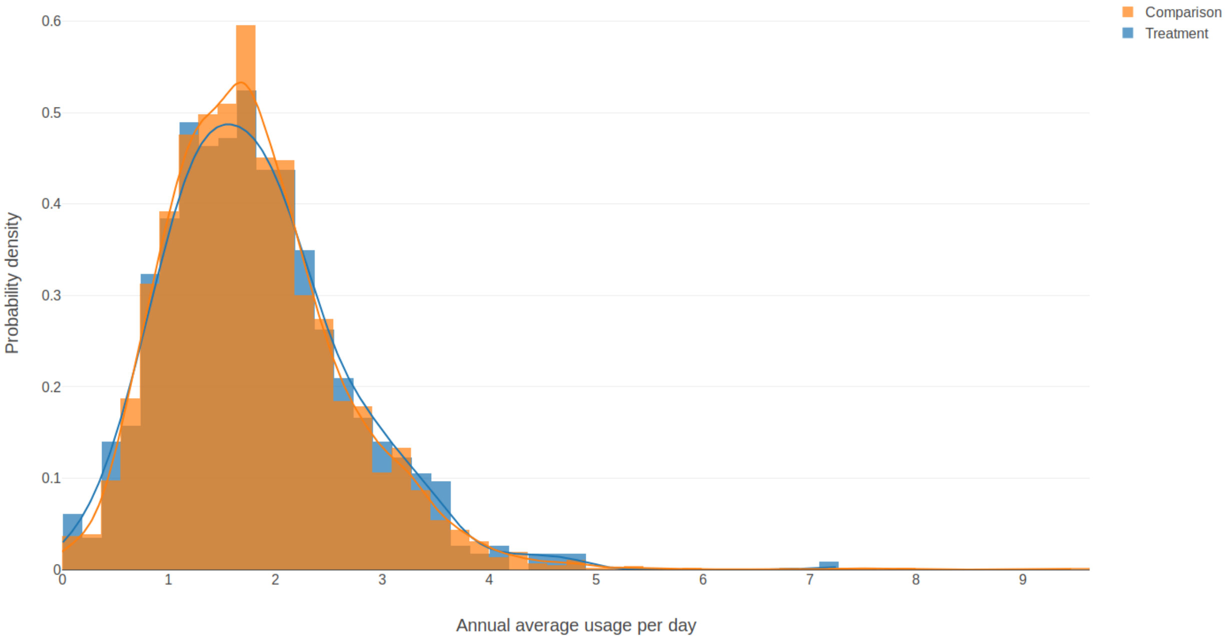


Figure A2b. Histograms of average annual electricity consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a future participant comparison group ($n_{treatment}$ =750; $n_{comparison}$ =584).

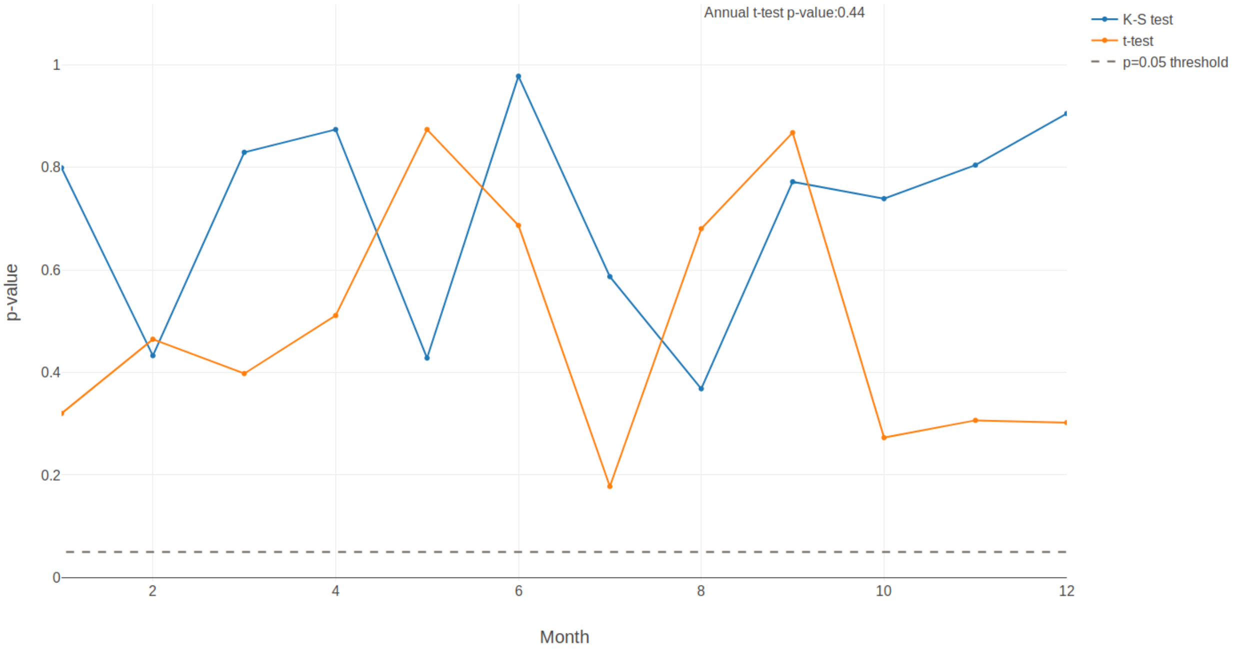


Figure A2c. Equivalence metrics for the treatment group (Ceiling insulation, 2013 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =682; $n_{comparison}$ =3367).

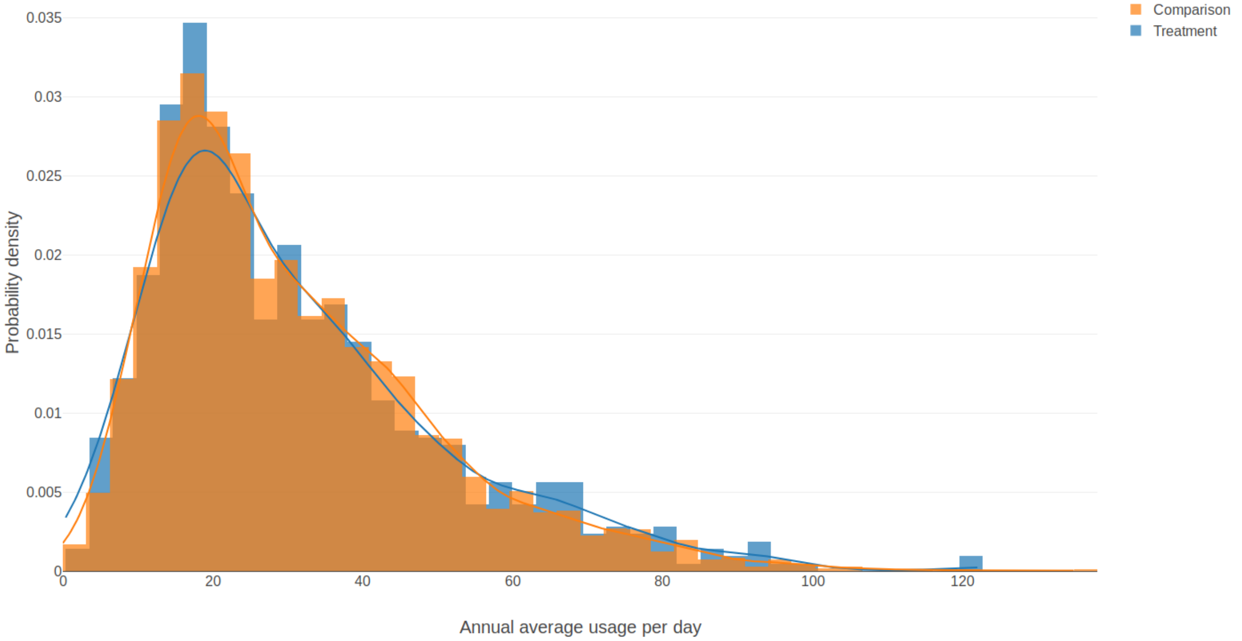


Figure A2d. Histograms of average annual electricity consumption per day for the treatment group (Ceiling insulation, 2013 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =682; $n_{comparison}$ =3367).

A3. Equivalence metrics for Ceiling insulation in 2014 (Electricity)

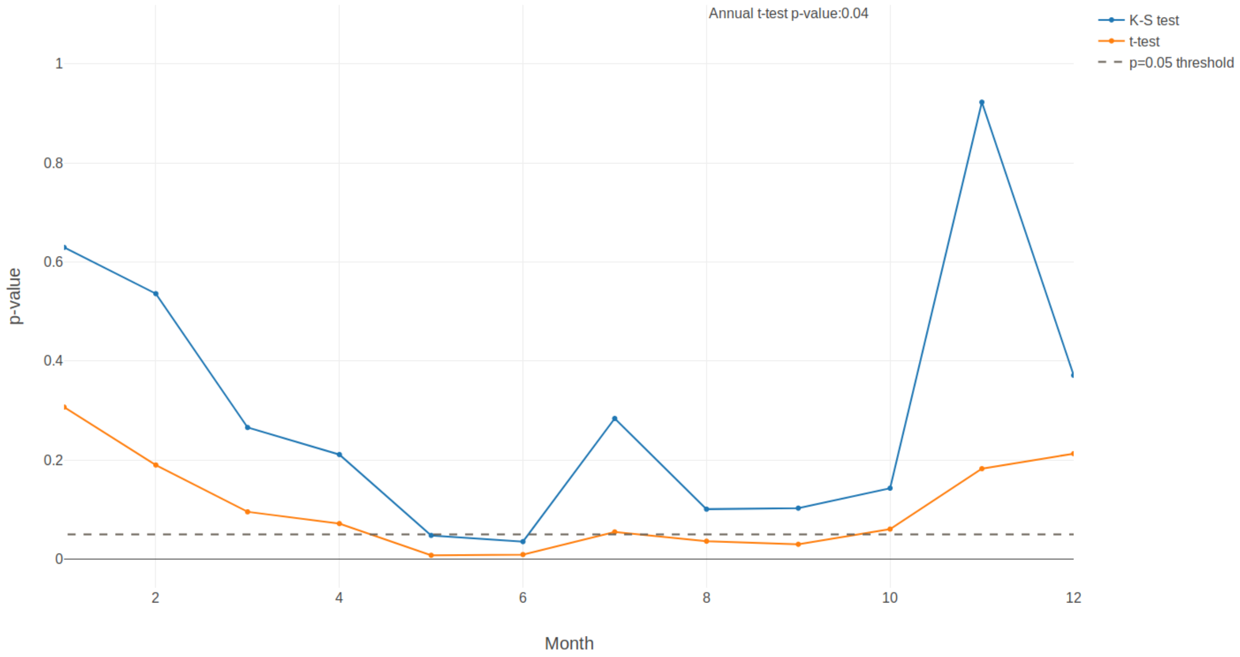


Figure A3a. Equivalence metrics for the treatment group (Ceiling insulation, 2014 program year) and a future participant comparison group (Ceiling insulation, 2015 program year) ($n_{treatment}$ =771; $n_{comparison}$ =733).

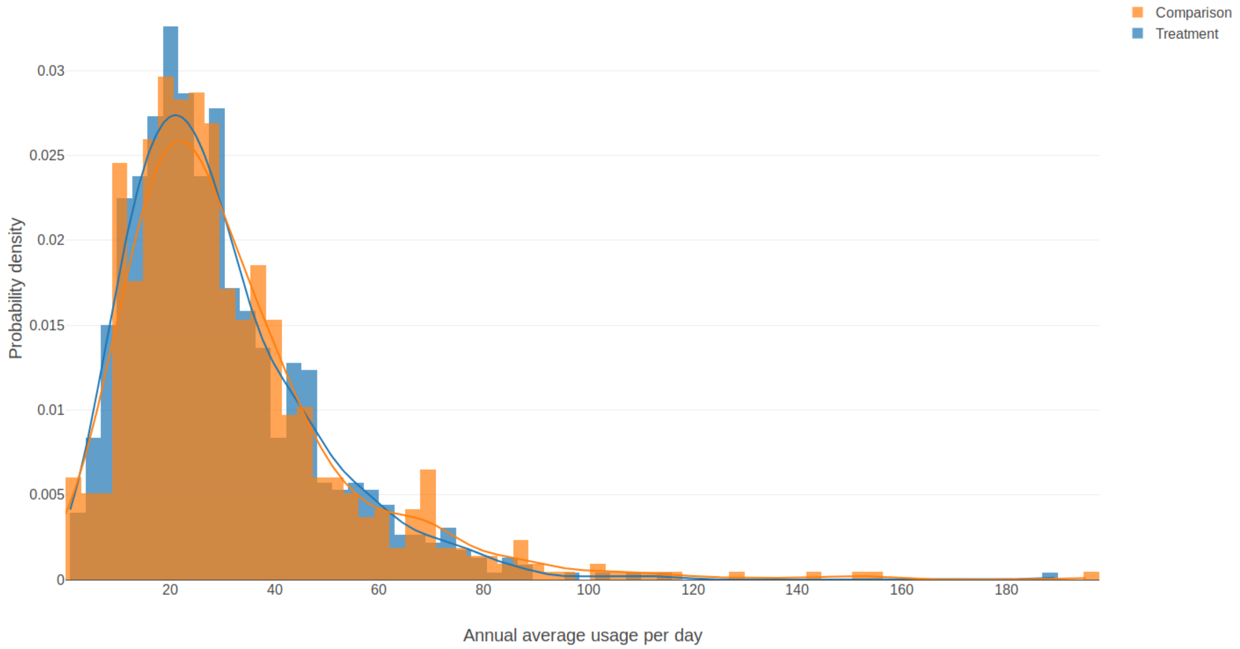


Figure A3b. Histograms of average annual electricity consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a future participant comparison group ($n_{treatment}$ =771; $n_{comparison}$ =733).

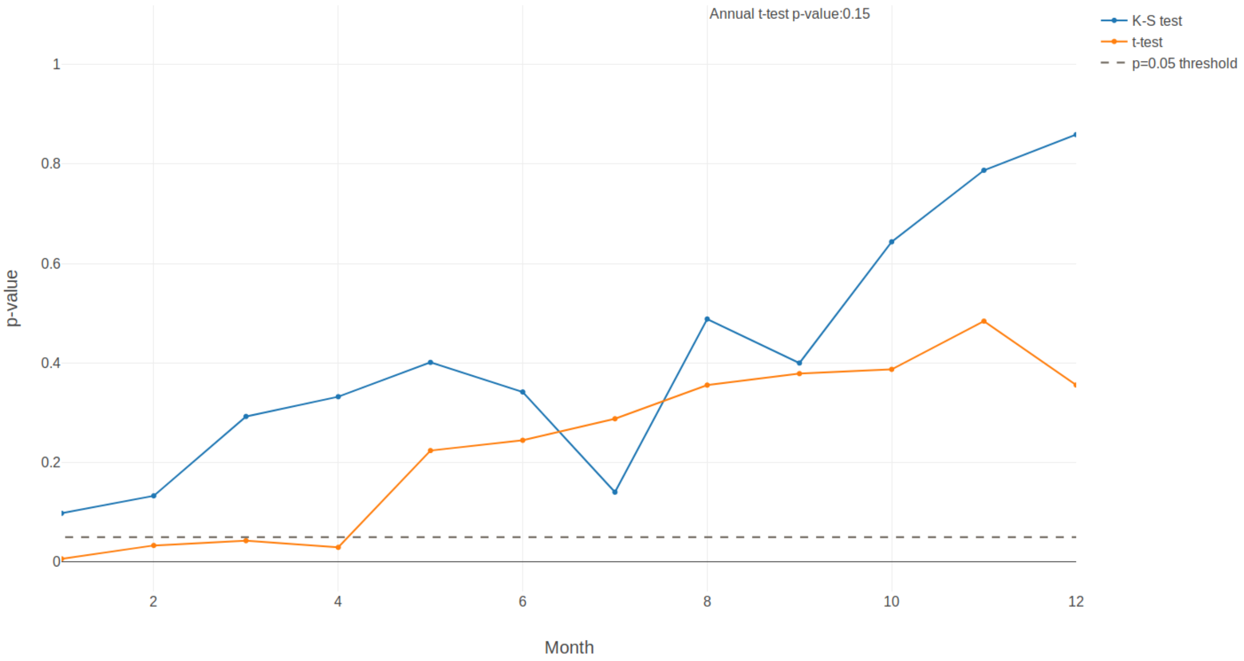


Figure A3c. Equivalence metrics for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =597; $n_{comparison}$ =2963).

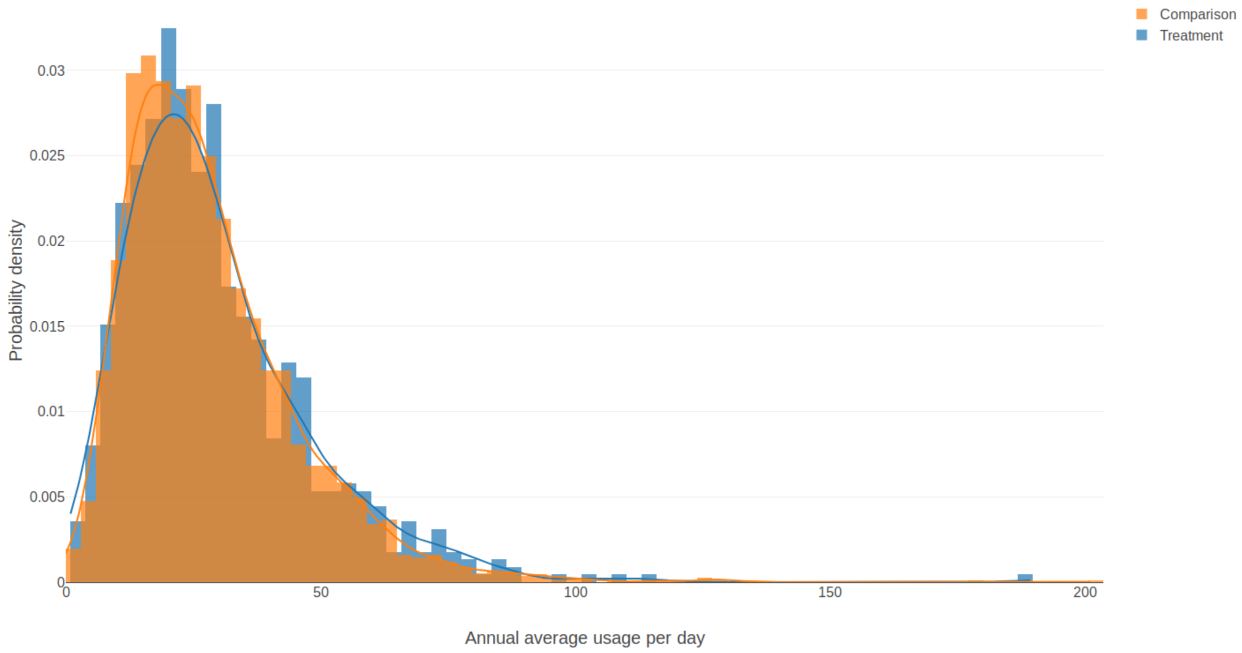


Figure A3d. Histograms of average annual electricity consumption per day for the treatment group (Ceiling insulation, 2014 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =597; $n_{comparison}$ =2963).

A4. Equivalence metrics for Ceiling insulation in 2015 (Gas)

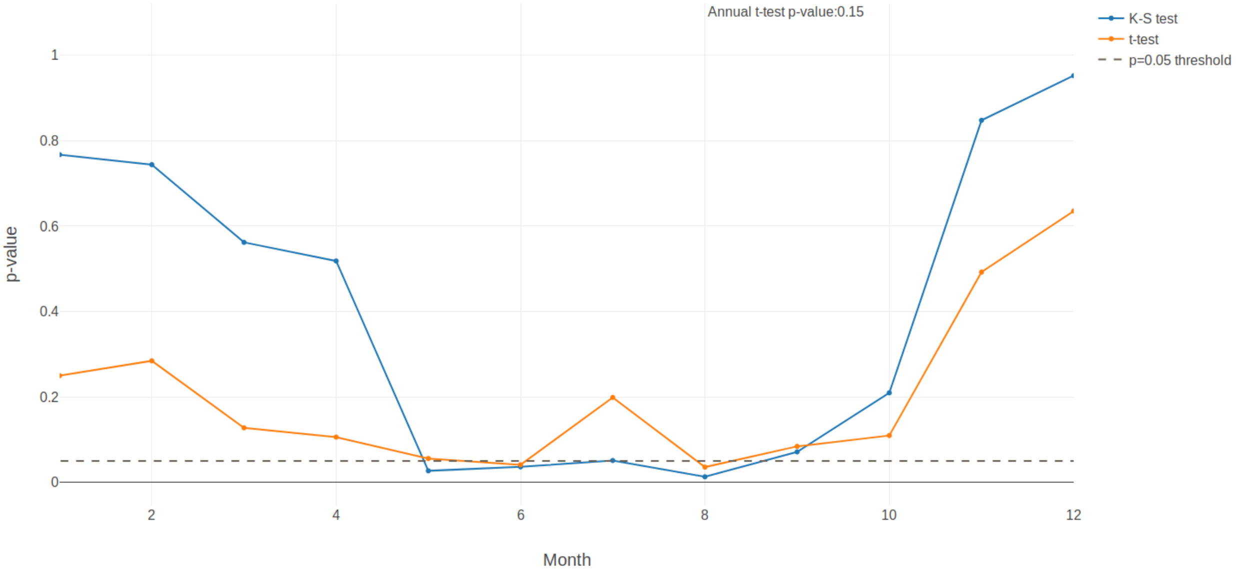


Figure A4a. Equivalence metrics for the treatment group (Ceiling insulation, 2015 program year) and a future participant comparison group (Ceiling insulation, 2016 program year) ($n_{treatment}$ =648; $n_{comparison}$ =478).

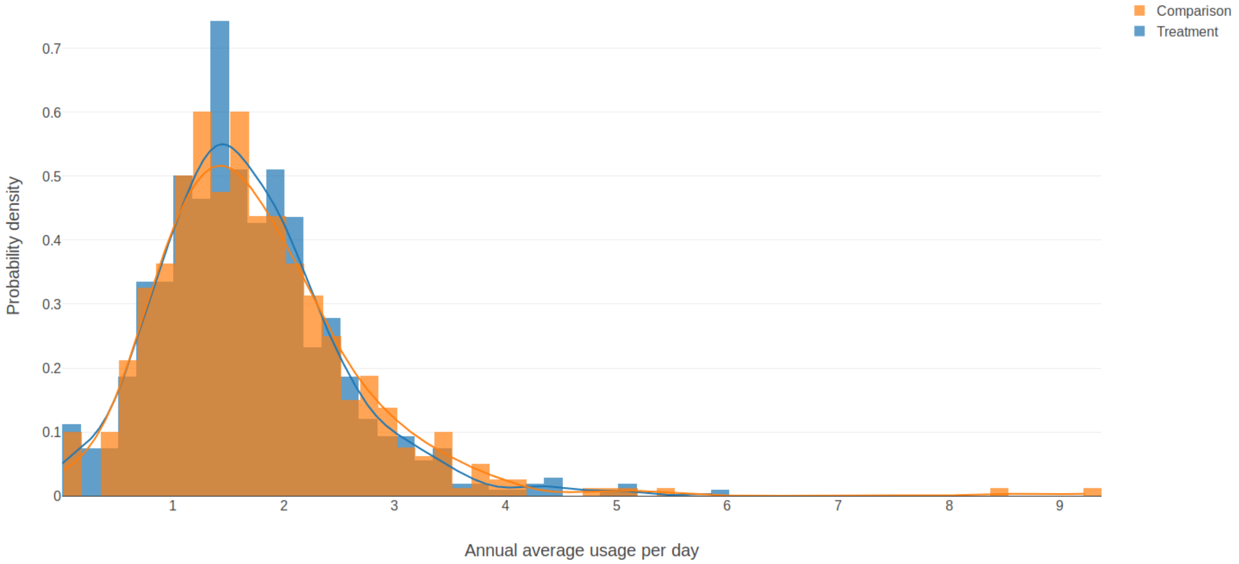


Figure A4b. Histograms of average annual gas consumption per day for the treatment group (Ceiling insulation, 2015 program year) and a future participant comparison group ($n_{treatment}$ =648; $n_{comparison}$ =478).

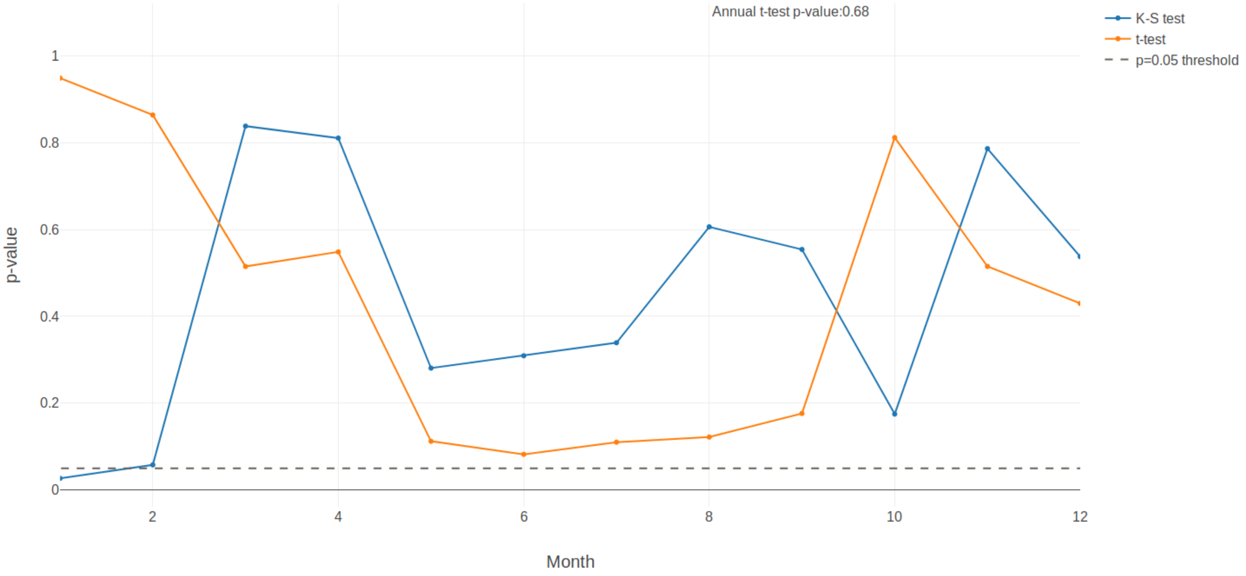


Figure A4c. Equivalence metrics for the treatment group (Ceiling insulation, 2015 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =644; $n_{comparison}$ =3200).

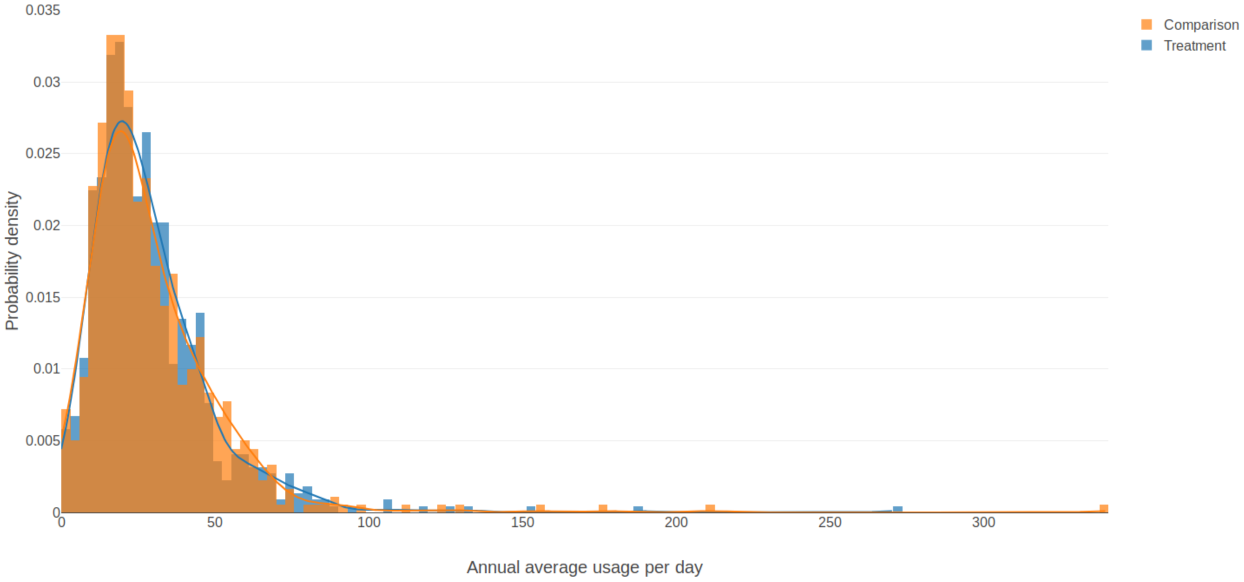


Figure A4d. Histograms of average annual gas consumption per day for the treatment group (Ceiling insulation, 2015 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =644; $n_{comparison}$ =3200).

A5. Equivalence metrics for Ceiling insulation in 2015 (Electricity)

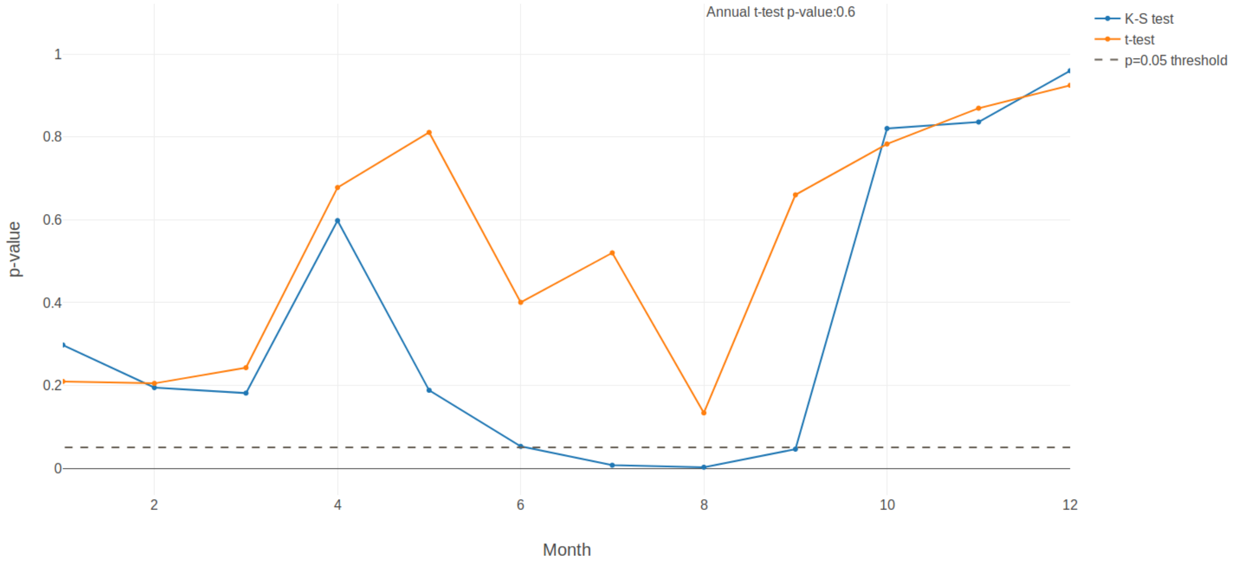


Figure A5a. Equivalence metrics for the treatment group (Ceiling insulation, 2015 program year) and a future participant comparison group (Ceiling insulation, 2016 program year) ($n_{treatment}$ =769; $n_{comparison}$ =620).

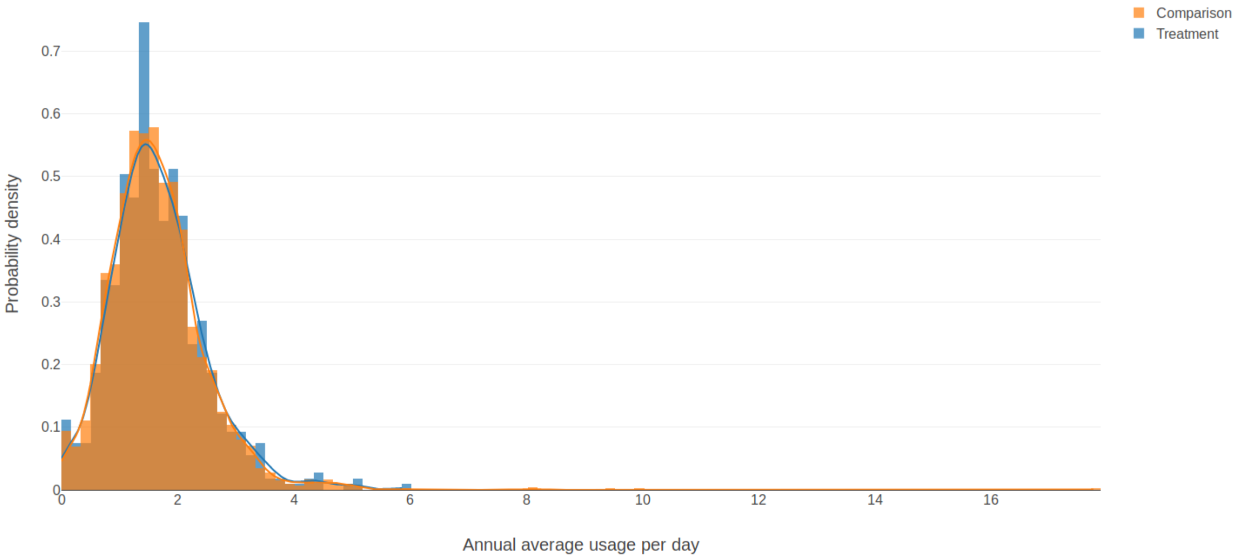


Figure A5b. Histograms of average annual electricity consumption per day for the treatment group (Ceiling insulation, 2015 program year) and a future participant comparison group ($n_{treatment}$ =769; $n_{comparison}$ =620).

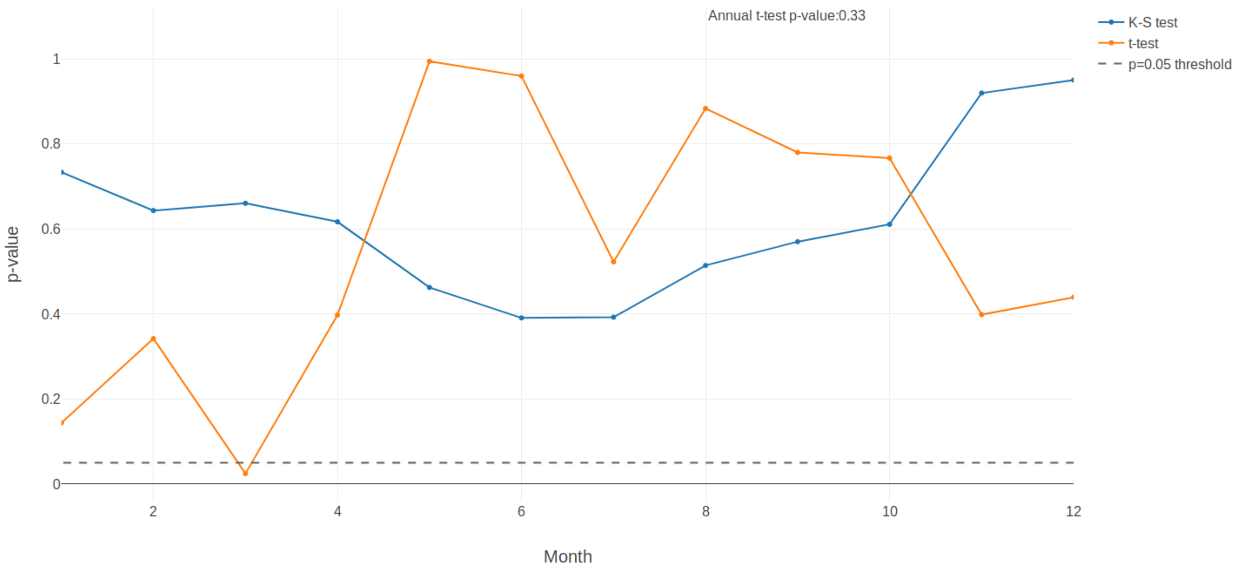


Figure A5c. Equivalence metrics for the treatment group (Ceiling insulation, 2015 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =766; $n_{comparison}$ =3788).

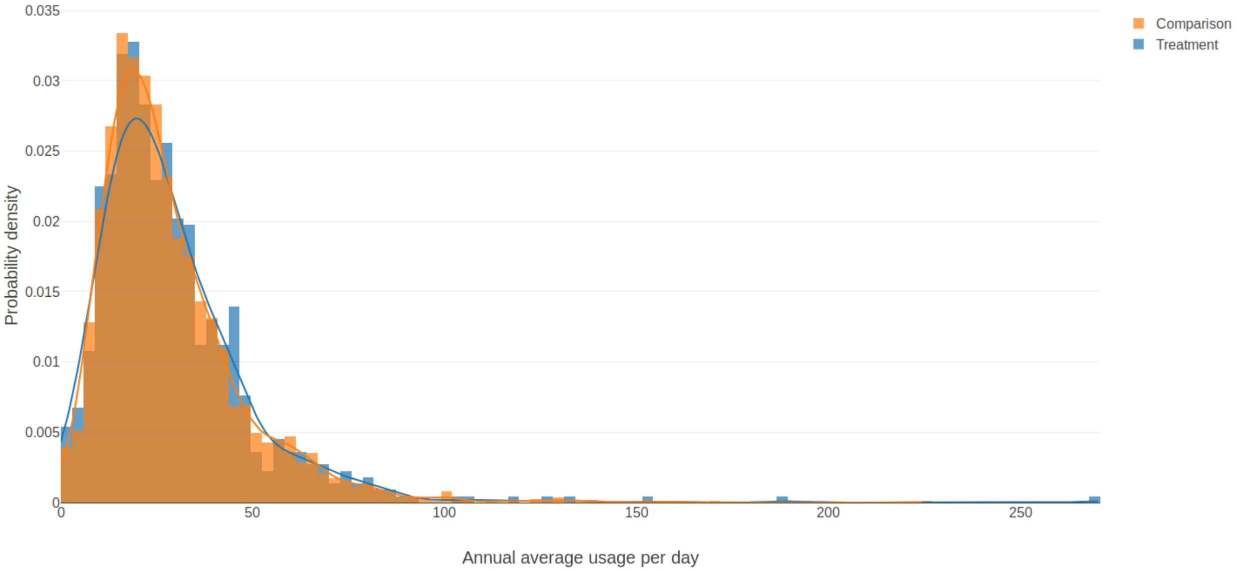


Figure A5d. Histograms of average annual electricity consumption per day for the treatment group (Ceiling insulation, 2015 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =766; $n_{comparison}$ =3788).

A6. Equivalence metrics for Smart thermostats in 2015 (Gas)

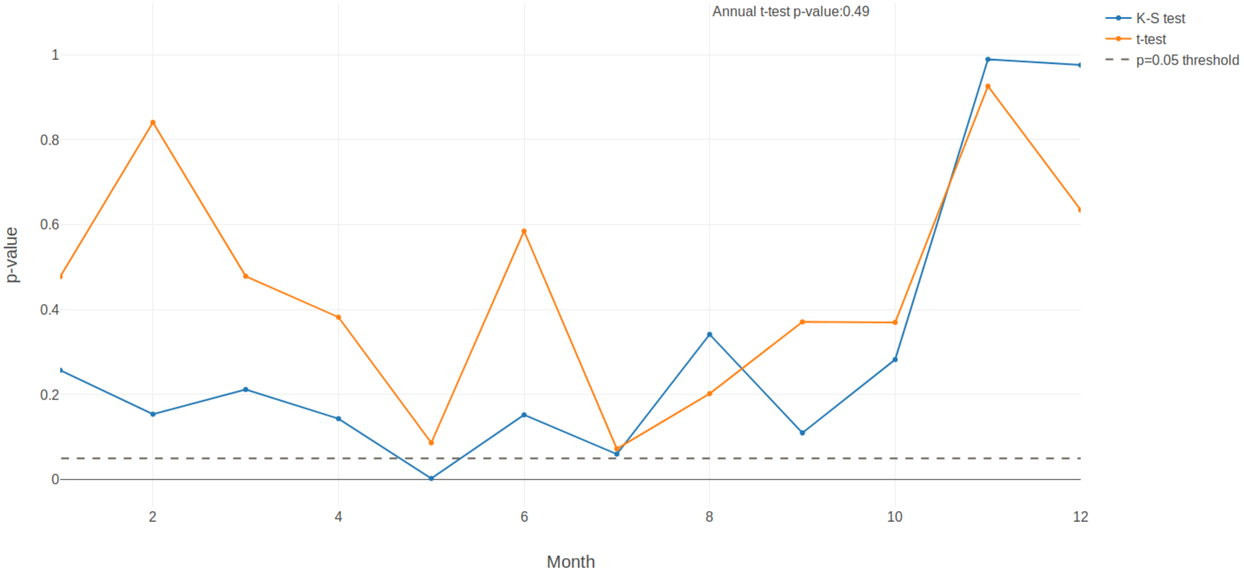


Figure A6a. Equivalence metrics for the treatment group (Smart thermostats, 2015 program year) and a future participant comparison group (Smart thermostats, 2016 program year) ($n_{treatment}$ =434; $n_{comparison}$ =1636).

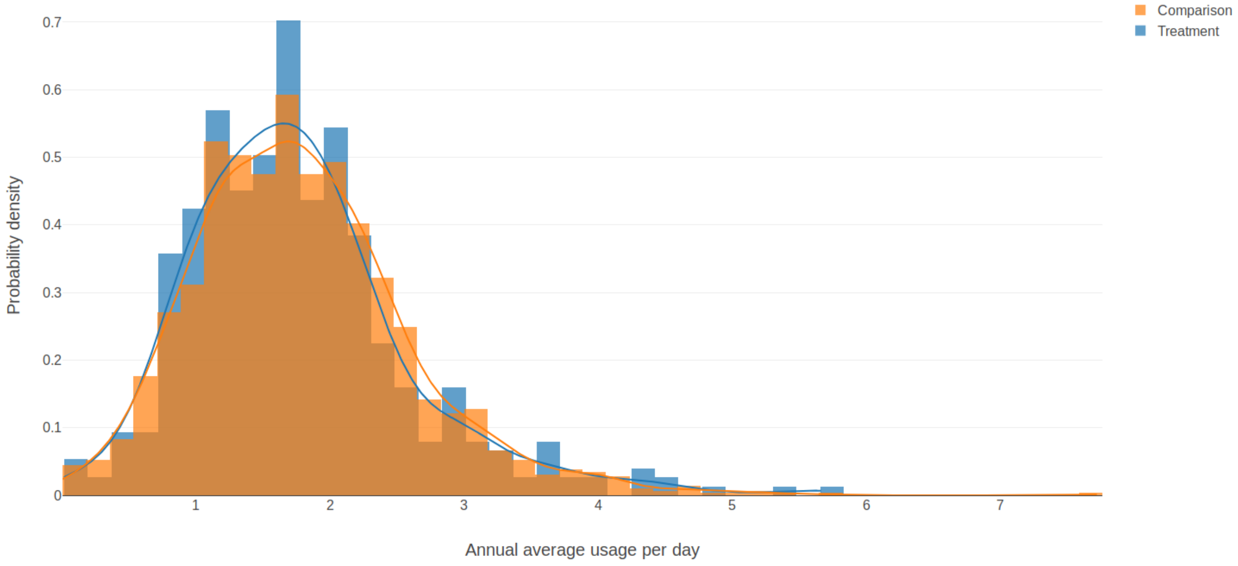


Figure A6b. Histograms of average annual gas consumption per day for the treatment group (Smart thermostats, 2015 program year) and a future participant comparison group ($n_{treatment}$ =434; $n_{comparison}$ =1636).

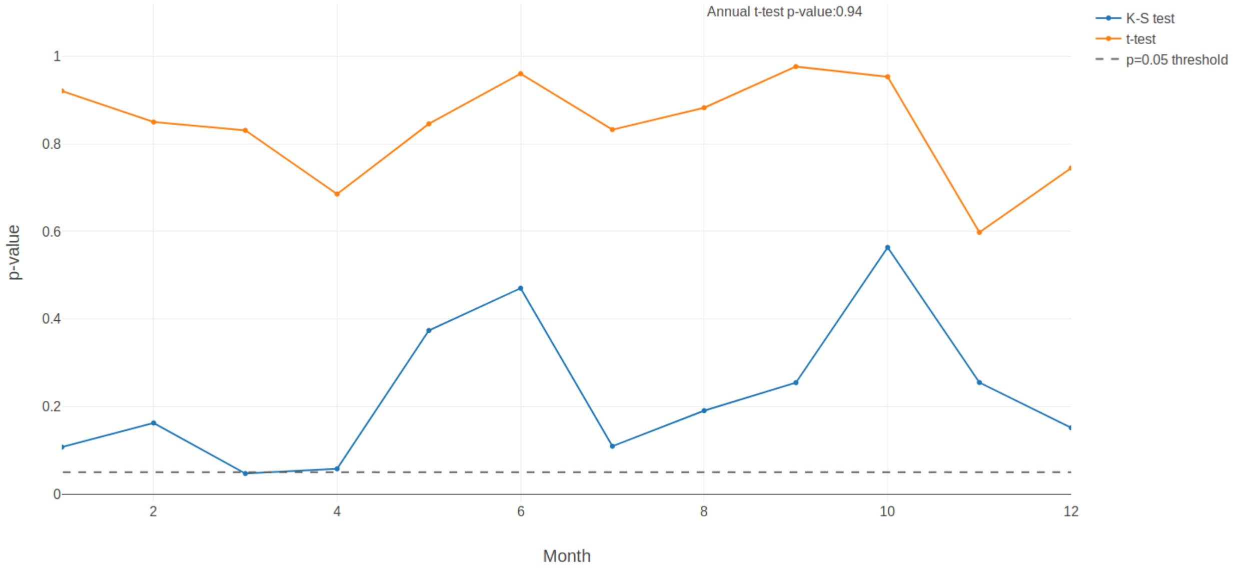


Figure A6c. Equivalence metrics for the treatment group (Smart thermostats, 2015 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =430; $n_{comparison}$ =2160).

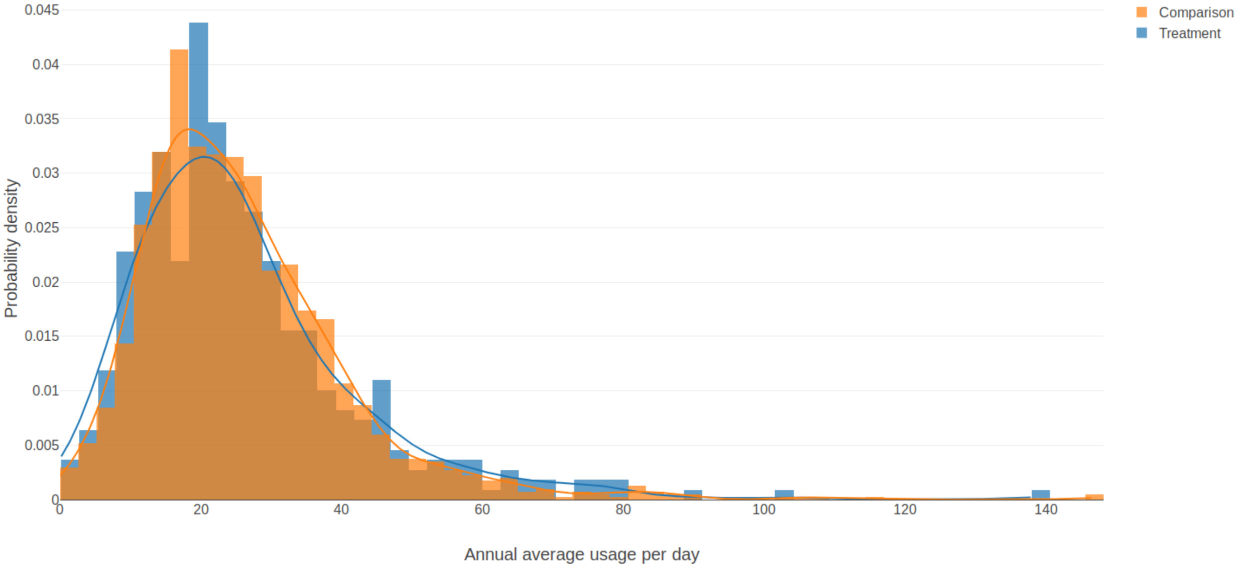


Figure A6d. Histograms of average annual gas consumption per day for the treatment group (Smart thermostats, 2015 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =430; $n_{comparison}$ =2160).

A7. Equivalence metrics for Smart thermostats in 2015 (Electricity)

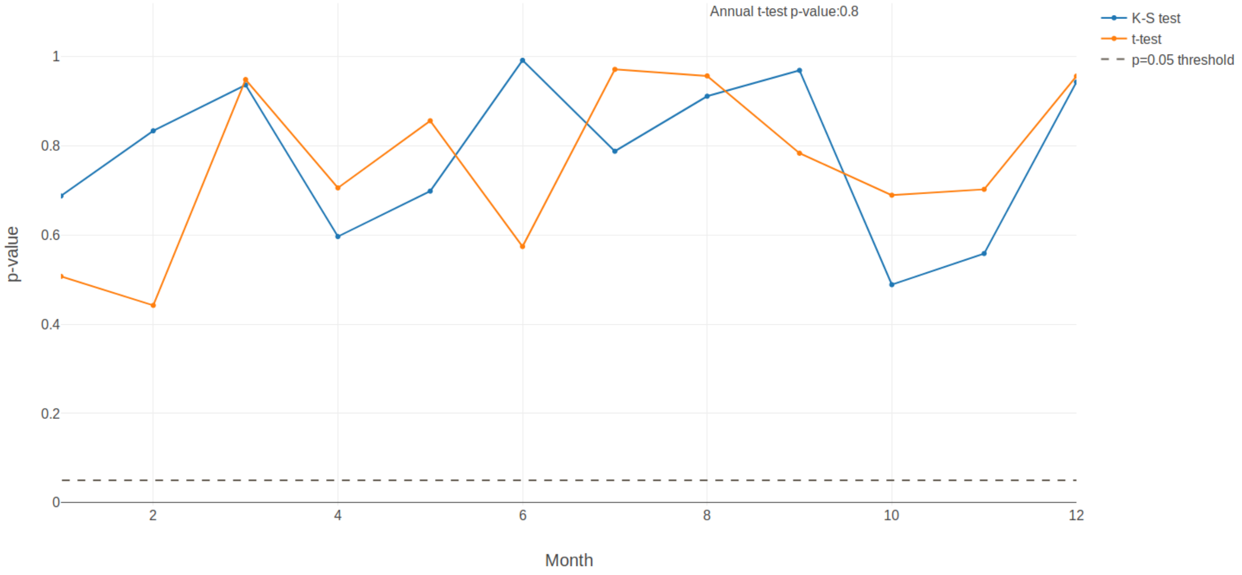


Figure A7a. Equivalence metrics for the treatment group (Smart thermostats, 2015 program year) and a future participant comparison group (Smart thermostats, 2016 program year) ($n_{treatment}$ =425; $n_{comparison}$ =1551).

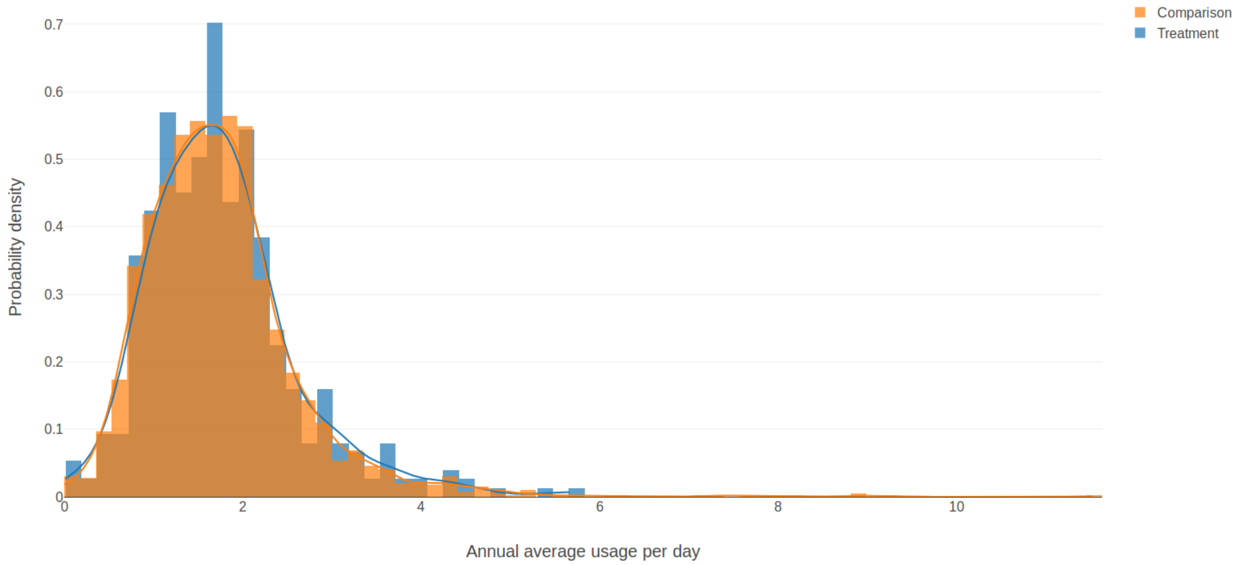


Figure A7b. Histograms of average annual electricity consumption per day for the treatment group (Smart thermostats, 2015 program year) and a future participant comparison group ($n_{treatment}$ =425; $n_{comparison}$ =1551).

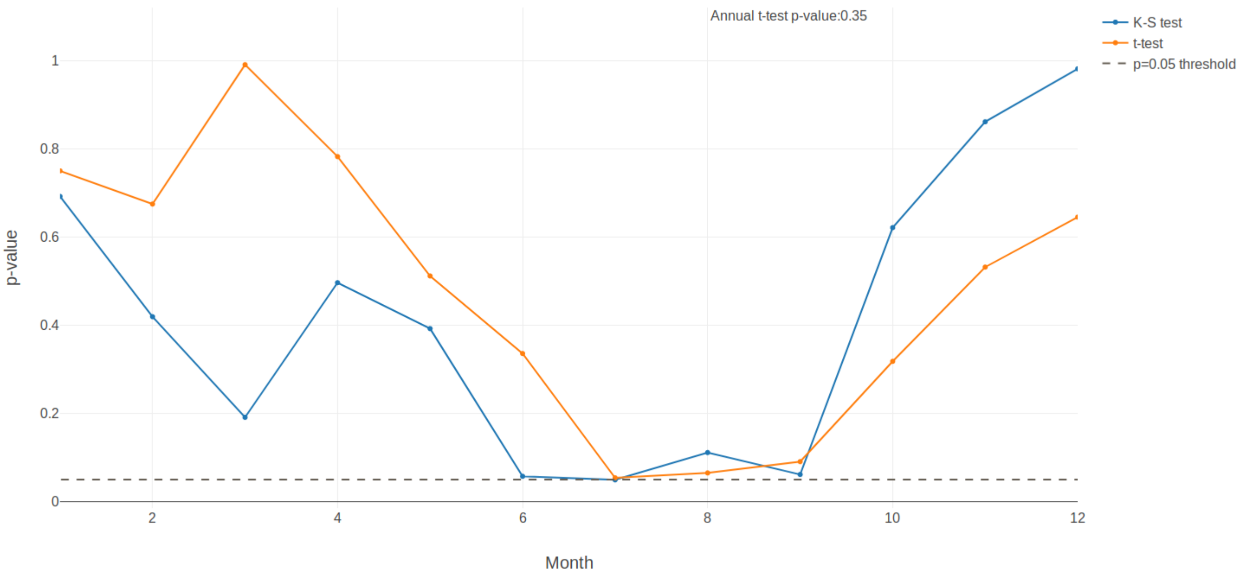


Figure A7c. Equivalence metrics for the treatment group (Smart thermostats, 2015 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =422; $n_{comparison}$ =2095).

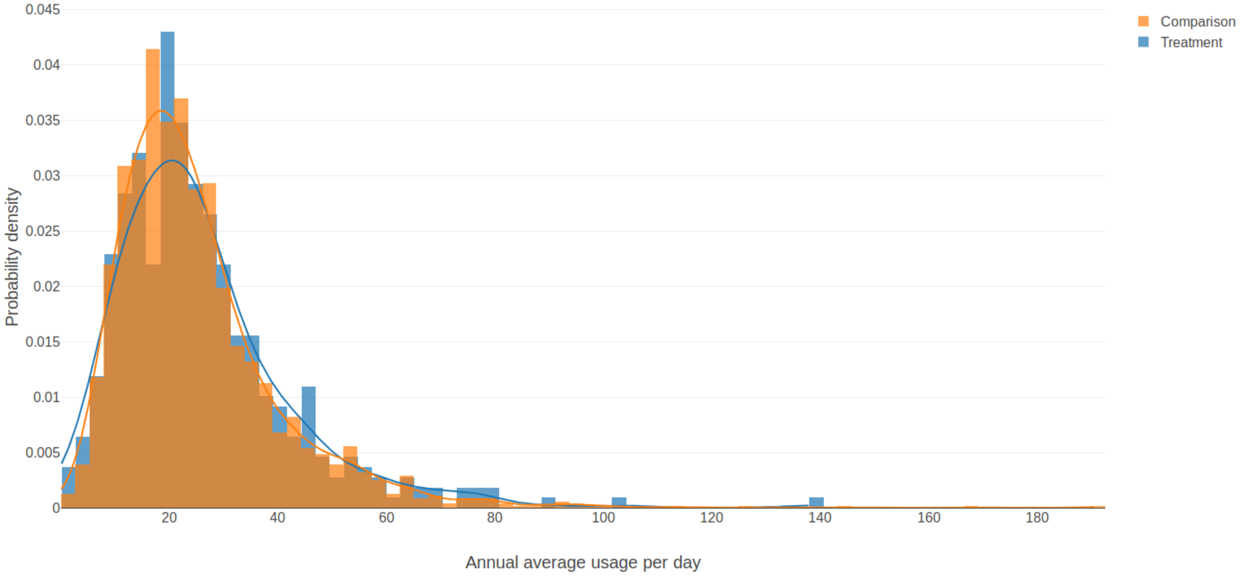


Figure A7d. Histograms of average annual electricity consumption per day for the treatment group (Smart thermostats, 2015 program year) and a comparison group determined by individual matching on monthly consumption ($n_{treatment}$ =422; $n_{comparison}$ =2095).