Control Group Wars - There's More Than One Way to Win the Battle

Dave Hanna, Itron, Inc., San Diego, CA
Kelly Marrin, EnerNOC, Inc., Riverside, CA

ABSTRACT

Broadly marketed utility sponsored behavior change programs are cropping up more frequently, especially at utilities that have completed Advanced Meter Infrastructure (AMI) rollouts. From social pressure driven programs like Opower, to energy dashboards that present efficiency tips and interval usage data, these programs seek to provide information to customers that will encourage them to change the way they think about and consume energy. Even in the best case, the energy reduction associated with behavioral programs is anticipated to be very small (1-5%) which makes evaluation challenging. In addition, some behavior-based programs are marketed to the general public without first conducting an experimentally designed pilot. In these imperfect situations, a quasi-experimental design with pre-post comparisons of treatment and control groups is gaining greater acceptance and more frequent use.

The paper describes two control group selection techniques used to evaluate two AMI enabled information feedback programs. Both evaluations employed a non-equivalent control group design which is structured similar to a pre-post randomized design (i.e., Randomized Control Trial) but it lacks the important quality of randomization. Because the control groups are not randomly selected, they can never completely eliminate self selection bias. However, they can be developed to control for bias from observable factors that influence energy usage and potentially savings. The application of these techniques for evaluating impacts goes far beyond AMI enabled conservation programs. They can be applied to the evaluation of community-based energy efficiency programs and opt-in dynamic pricing or demand response programs.

Introduction

As more utilities move their behavior programs from the pilot phase into full deployment it becomes impossible to use true randomized experimental methods to create control groups for evaluation. In the absence of a randomized control group, a quasi-experimental design is often used to create a control group where a decent-sized pool of non-participating customers exists from which matched controls can be selected. The goal of creating a matched control group is to create a group of non-participants that looks as much like program participants as possible. There are at least three methods for creating matched control groups being used in the industry today, stratified matching, propensity score matching, and Euclidean distance matching. Both Itron and EnerNOC recently evaluated similar utility information programs using a quasi-experimental design. Each firm created the matched control groups using different methods, and both methods yielded good matches. The goal of this paper is to examine the two different methods used, to discuss the pros and cons of each method, and to offer some general advice for those working with matched control groups based on our lessons learned.

Two Methods for Creating Matched Control Groups

The control group matching methods discussed in this paper are Stratified Propensity Score Matching (SPSM) used by Itron and Stratified Euclidian Distance Matching (SEDM) used by EnerNOC.

Data Development

The programs that are used as case studies for discussing the control group selection are utility-sponsored behavior-based energy conservation programs that leverage interval load data collected by smart meters as part of two full-scale deployments of these meters. These behavior based programs can be characterized as web presentment and bill alert programs that are made available to virtually all the residential customers at these two utilities. The data that are available for all residential customers are primarily electricity usage data.

The data used in the development of the control groups was either related to the energy usage of the customers or was more of a categorical nature, e.g. geographic location. The development of the energy usage variables involved using the either monthly billing cycle usage and or hourly interval usage. Energy usage variables developed included actual monthly consumption, weather sensitivity of usage, summer to winter ratio of consumption, ratio of total usage to summer usage, and the coefficient of variation of monthly consumption. The categorical data included meter districts, zip codes, and rate classifications (i.e. income assistance.) The categorical data was used primarily for stratifying the participant and non-participant populations.

There are issues to consider with using either interval or monthly billing cycle usage data. Before developing the energy consumption related variables for the different methods, one needs to consider the pre-processing that is required for either form of usage data. Monthly billing data will not be the same for all customers as each customer's meter is typically read on one of several cycles; approximately one for each weekday in the month. This means that the data must be calendarized. Calendarization involves apportioning the usage to specific calendar months based on the number of days the cycle month falls within the actual calendar months it straddles. The result is an approximation of the usage in the calendar months and not the actual usage. If the usage is highly weather sensitive and there are significant differences in the weather between two adjacent months, the approximation may not be very accurate.

With the use of hourly interval usage data, there is the consideration of the volume of data that must be processed. If the pool of participants and especially the non-participants is very large there may be a significant amount of data to process in order to develop the required data for selecting a control group. Neither of the evaluations discussed in this paper examined the effect of the improved accuracy of the usage data on the outcome of the analyses as a result of using hourly interval usage. One can aggregate interval data by calendar month instead of using calendarized billing data.

As with all analyses, data cleaning was a necessity. The cleaning processes may be better described as data screening and elimination to provide a clean dataset for analysis purposes. The cleaning processes used depended partially on the type of usage data (interval vs. monthly cycle usage). Interval usage data needs to be screened for completeness of all hours and all days. The monthly cycle usage data needed to be screened for completeness as well (i.e., all months.) Another consideration in the cleaning process is whether or not a given customer is truly an eligible customer. Not all residential electric accounts are what they seem. Very small residential accounts can be associated with something other than a residence. Very large accounts can represent master metered multi-family buildings where the tenants don't pay the bill and therefore aren't eligible to participant in the behavioral programs. It is necessary to make judgments about what constitutes a true potential participant in terms of the level of energy consumption. The program sponsoring utility may be able to provide guidance on this. Large and small usage accounts were removed from the pool of non-participants in both evaluations and treated as outliers as they deviate significantly from the remaining customers. Leaving these accounts in the analysis can mask what is anticipated to be a very small effect to begin with.

The data cleaning process also eliminated sites from the participant and nonparticipant population that were participants in the California Solar Initiative (CSI). These sites were eliminated from the analysis dataset due to the difficulty in finding a matched nonparticipant sample among so few sites and because these sites are likely to have lower than average levels of utility energy consumption.

The low levels of utility energy consumption would make it more difficult to determine the observable potential impact of the program on their billing data.

Method A: Stratified Propensity Score Matching (SPSM)

As mentioned previously, the energy impacts of certain behavioral programs must be estimated using quasi-experimental matching methods. The matching methodology uses observable characteristics of the participant and nonparticipant groups to develop groups that are similar in all observable characteristics. The observable customer characteristics for the first matching method discussed here include CARE[1] status, household geographic location, usage, and the seasonal distribution of usage. Matching on these observable characteristics would likely reduce the potential bias in the estimate of the program's impact. The reduction in bias is due to the assumption that households with similar observable characteristics also have similar views on carbon footprints.

Propensity score matching with stratification (SPSM) was the matching methodology used for the evaluation of SCE's My Account and Budget Assistant programs. The propensity score represents the probability of participation based on pre-program period observable characteristics. Most of the characteristics used to assign this score were derived from kWh recorded during the 12-month period February 2010 to January 2011. The SPSM methodology uses observable classification variables to stratify the participant and nonparticipant populations and then use a logistic regression (logit) model to estimate the probability of participation within the participant and nonparticipant strata.

Stratification. Stratification of the populations prior to scoring controls for variables that may observationally influence participation. To guarantee that the participant and control groups were balanced in terms of geographic location and CARE status, the sites were first stratified by these variables. Geographic location influences participation due to the fact that AMI meter installations were undertaken based upon location. Geographic location may also influence participation based on unobservable factors such as the average computer and internet usage rates of the region. The combination of CARE and geographic location led to the development of six strata within the propensity scoring methodology.

Model Development. Within each of the six strata, a logit model was used to estimate the probability of program participation. Using logistic regression, propensity scores were first estimated for all Budget Assistant participants and nonparticipants (i.e., customers who were in the participant/nonparticipant sample but not Budget Assistant or My Account participants). This propensity score represents the probability of participating in the Budget Assistant program. Most of the characteristics used to assign this score were derived from kWh recorded during the 12-month period February 2010 to January 2011.

Within each of the six strata, a logit model was used to estimate the probability of program participation.

The independent variables used in the logistic regressions include:

Corr_summer: Customer-level Spearman Rank-Order correlation coefficient correlating monthly CDD65 and kWh usage over the months April-November 2010.

Corr_winter: Customer-level Spearman Rank-Order correlation coefficient correlating monthly HDD65 and monthly kWh usage[2] over the months February-May 2010, October-December 2010, and January 2011.

Cov: Coefficient of variation, February 2010 to January 2011.

[1] CARE stands for California Alternative Rates for Energy and is a program whereby low income customers may receive lower rates than non-CARE customers.

[2] The billing and weather data for all customers in the participant/nonparticipant population were "calendarized" so that monthly values could be compared across customers.

KWh_1, …, KWh_12: Monthly kWh usage, February 2010 to January 2011.
KWh_2009_7, …, KWh_2009_9: Monthly kWh usage, July 2009 to September 2009.
KWh_avg_mo: The average monthly kWh usage over the months February 2010 to January 2011.
Ratio_hot_to_cold: The ratio of average monthly usage of July-September 2010 to November-April 2010. The purpose of this variable was to make the logistic model more sensitive to customers who increase their usage during the summer months.
Ratio_u_to_cdd: The ratio of total kWh usage to total CCD65, over the summer months of 2010. The purpose of this variable was also to make the logistic model more sensitive to customers who increase their usage when temperatures increase.

Matching Process. Nonparticipants with similar scores to the Budget Assistant participants were selected as nonparticipant matches and removed from the nonparticipant sample available for My Account matching. Next, a second propensity score was assigned to the My Account participants and remaining nonparticipants. This score represented the probability of participating in My Account. The matching process for My Account participants was conducted based on these scores estimated with the logit models.

The mean values of all covariates used to assign the propensity scores for participants and nonparticipants were compared both before and after the match. **Error! Reference source not found.** contains a subset of the matching variables for all Budget Assistant participants, nonparticipants, and Budget Assistant-matched nonparticipants. The statistical significance of the difference in means t-test is shown in the column next to each nonparticipant column. The difference in means testing compared the participant mean value with the respective nonparticipant's mean value. The *Probability of Program Participation* is the average propensity score for the group.

Table 1: Budget Assistant Covariate Balance

Variable	**Participant Mean**	**Nonparticipant Mean**	**Nonparticipant t-test (p value)**	**Matched Nonparticipant Mean**	**Matched Nonparticipant t-test (p value)**
corr_summer	0.67	0.68	0.5006	0.68	0.5541
corr_winter	0.10	0.15	0.0028	0.11	0.7524
cov	19.99	22.88	<.0001	19.76	0.6606
kwh_2009_7	933	878	0.0032	947	0.6248
kwh_2009_8	954	883	0.0002	962	0.7924
kwh_2009_9	925	849	<.0001	931	0.8251
ratio_u_to_cdd	4.89	3.29	<.0001	4.88	0.9530
Probability of Participation	0.00948	0.00416	<.0001	0.00932	0.7309
Num Service Accounts	832	198,181	n/a	832	n/a

Reviewing the statistics in **Error! Reference source not found.**, the average monthly usage for participants in Budget Assistant for the period February 2010 – January 2011 was 680 kWh. This value is significantly higher than the average monthly usage for nonparticipant, 622 kWh. Following the matching process, the average monthly usage for the matched nonparticipants was 677 kWh, a value not statistically different from the average consumption of participants. As shown in **Error! Reference source not found.**, following the matching process the independent variables used in the participant scoring model are similar for the participant and nonparticipant samples.

The matching process was repeated for the My Account participants. The average monthly usage of My Account participants was 671 kWh. The usage of participants is statistically significantly higher than the 621 kWh average monthly usage of nonparticipant. Following the matching process, the average monthly usage for matched nonparticipant was 668 kWh.

Results. Figure 1 and Figure 2 present the monthly load profiles for the Budget Assistant participants and the population of nonparticipants prior to matching and following matching respectively. Prior to matching, the load profile for the nonparticipants is always substantially lower than the profile for the Budget Assistant participants. Following the matching process, the load profiles for the participants and the matched nonparticipants are largely indistinguishable.

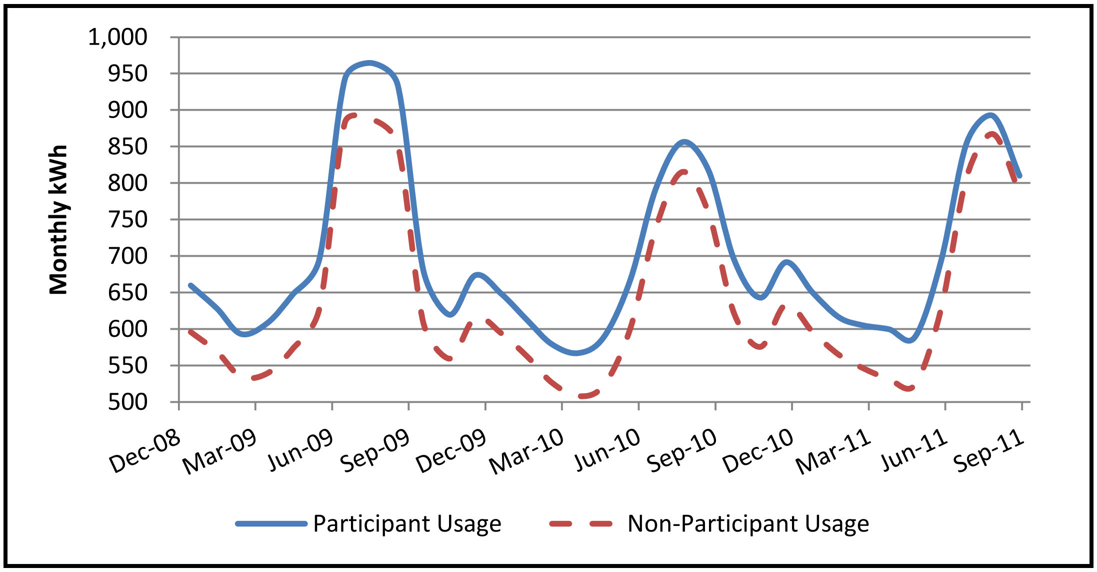

Figure 1: Budget Assistant Monthly Load Profiles – Pre-Match

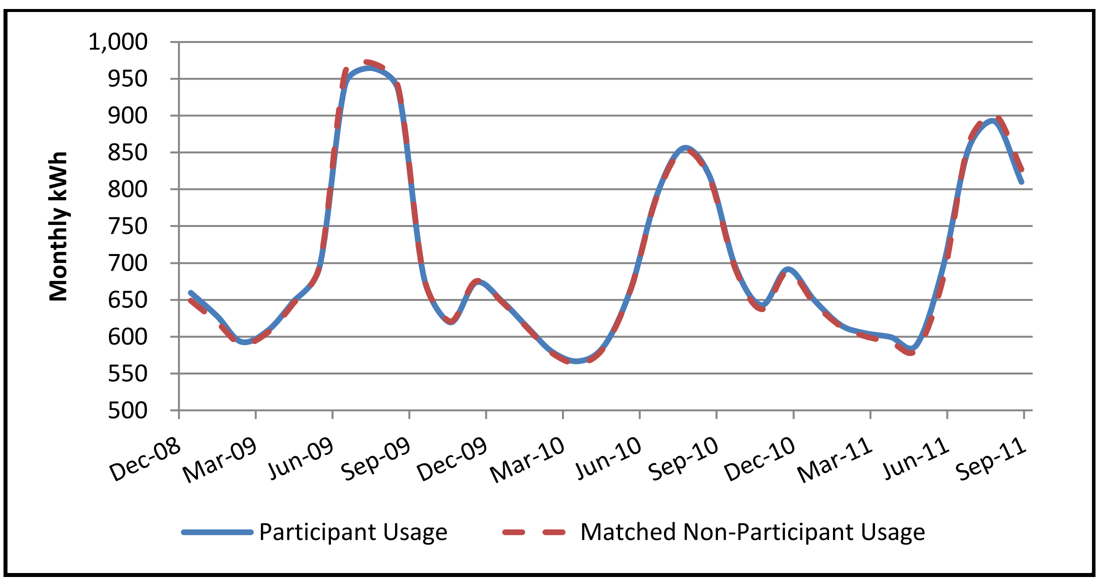

Figure 2: Budget Assistant Monthly Load Profiles – Matched

To verify that the selection process was not introducing any additional bias in the larger analysis of energy savings, Itron conducted a Monte Carlo analysis to ensure that the results from the billing analysis were not an artifact of the specific non-participant sample chosen during the matching process, that the results are robust to random modifications to the non-participant control group. Given that each participant was matched with just one of many candidate non-participants from a large sample frame, the SmartConnect billing analysis results could be influenced by some bias in the control group analyzed for the final analysis.

For this Monte Carlo analysis, Itron estimated 500 billing analysis models comparing the consumption of participants and control groups of non-participants. The non-participant control groups were selected using the SPSM method with random number assignment where the seed of the random number generator was varied for each iteration to develop different control groups for each of the 500 models. The variation of the random number generator seed means that the participant and non-participant matches within each iteration will result in substantially different control groups across iterations. The next step was to verify that the iterations of the site matching routine resulted in sufficiently different sets of matched non-participants. If the matching produced control groups that all consisted of similar sets of accounts, then there would be no point in running through the regression portion of the analysis, as the results would certainly not vary from the original analysis. The verification step attempted to maximize the number of unique non-participants across all control groups. Using this approach it was possible to determine how sensitive the estimates of the model are to the non-participant sample.

To illustrate the results of the Monte Carlo analysis, plots of the distributions of t values from the analysis were developed. The range of t values produced by the analysis speak to weather it would have been possible to find a parameter statistically significant by choosing a different non-participant group. For example, the final report found that there were only two program parameters that were statistically significant. The findings presented in Figure 3 are consistent with the findings presented in the final report. The program parameter BANOTI_10 (Budget Assistant notifications equal to 10 or more) had t statistics that are consistently larger in absolute value than 1.96.

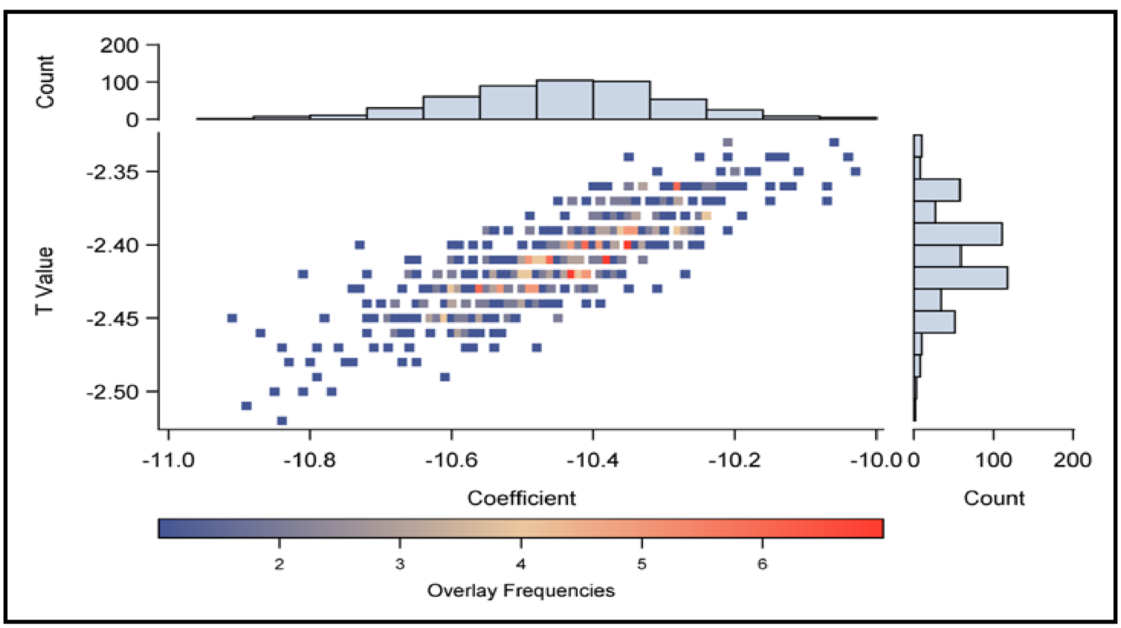

Figure 3: Scatter of Estimate by Robust T Value for BANOTI_10

Method B: Stratified Euclidian Distance Matching

The second method that is discussed in this paper has been used by EnerNOC to evaluate several utility Smart Meter enabled conservation programs two of which are Pacific Gas & Electric's comparable smart meter enabled programs, Customer Web Presentment (CWP) and Energy Alerts. The first step in creating a matched control group for any participant group is to define the non-participant population. This almost always results in the exclusion of specific customers, such as those participating in other utility programs, or specifically selecting from a group of non-participants that are thought to be similar to participants in some way, for example customers known to have internet access. In this case, the non-participant population for the Energy Alerts program was drawn from the general population, however the CWP participants were matched to those customers that had My Energy accounts but had not viewed their interval data. This ensured that both the CWP participants and controls would have access to the same web tools from the My Energy website.

Stratification. Similar to a stratified propensity method, once the non-participant population is identified both the treatment and candidate control group pools are assigned to strata or filters that are categorical in nature. For the 2011 program year more refined filters were used including 5 digit zip code, presence of air conditioning, electric heat, house type, CARE status, and participation in other DR programs. For the 2012 program year less refined filters were used including climate region, and house type (single vs. multi-family). In either case, the purpose of stratifying or filtering customers based on non-energy characteristics is to try to capture attributes that might affect the way customers use energy and the way they respond to the program. These characteristics are not directly observable, but are likely to be correlated with the characteristics that we can observe.

Once the filters are established all of the customers, both treatment and control are assigned to a specific group based on their strata. For example, for the simple stratification employing only climate zone and housing type the individual groups were: inland single family; inland multi-family; coastal single family; and coastal multi-family. At this stage, it is critical to ensure that there are enough candidate control group customers in each group. Usually a ratio of 10 control customers to each treatment customer is sufficient; however even larger ratios can yield a better match.

Euclidean Distance Metric Matching. After separating the treatment and potential control group customers into groups based on categorical stratification variables, the next step is to match each treatment customer with a similar control group customer from the same group based on pre-treatment electricity usage.

In order to determine how close each treatment customer is to a potential match we calculate a Euclidean distance metric. The Euclidean distance is defined as the square root of the sum of the squared differences between the matching variables. Any number of relevant variables could be included in the Euclidean distance, in the most recent evaluation we included twelve months of pre-treatment calendarized billing data with a rolling pre-treatment window that was based on the participant's start date. The Euclidean distance for this set of variables can be calculated by Equation 1 below.

(1)

In this example, the Euclidean distance in Equation 1 above would be calculated for participants within each enrollment window as the distance between each treatment customer and all potential candidate control customers. Using a distance metric allows us to compare treatment customers with potential control customers based on their overall similarity as defined by the pre-treatment monthly usage variables that are included in the Euclidean distance. It is also possible to weight the variables that are being included in the distance metric. For example, if summer is more important than winter, the summer months can be given more weight and the winter months can be given less weight.

After calculating the distance metric within each group for each possible combination of treatment and control customer, the control customer with the smallest distance is matched to each treatment customer in question without replacement. When we minimize the distance metric, we are in essence selecting the control customer who is the most like the treatment customer, in comparison to all other potential controls, across all of the variables included in the Euclidean distance. The process is repeated until all the treatment customers in the group have a match.

Results. After selecting a match for each control group customer, we can then examine the results of the match in the aggregate to determine how well the process worked. Figure 4 shows pre-treatment monthly usage for the Energy Alerts treatment and control group in program year 2011. In this example a more complex filtering method was used that included 5 digit zip code and several additional demographic variables creating very small groups of treatment customers and a potential control group pool with approximately a 10 to 1 ratio of control to treatment customers. In the early part of the year the match is very close, however as we approach the end of the year, the two groups begin to deviate somewhat with the treatment group always being slightly higher than the control group.

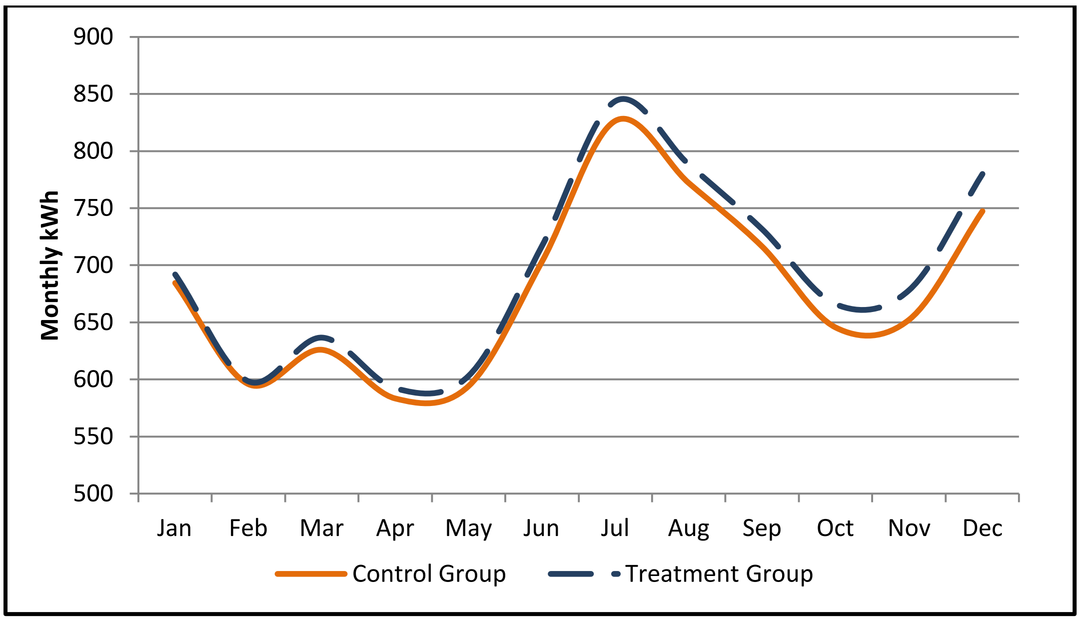

Figure 4: Energy Alerts Monthly Load Profiles - PY 2011 Match

Alternatively, in Figure 5 we used a less refined pre-stratification that included only two filters, creating four total groups. In addition we did not limit the control group pool in any way and selected from all eligible non-participants, thus significantly increasing the ratio of control to treatment to somewhere on the order of 725 to 1. In Figure 5 the match is significantly improved with both groups having nearly identical monthly usage over course of the year. The very refined groups used in 2011 created very small groups of customers in which to match, with between 5 and 20 customers in each group. Whereas creating larger groups with a few hundred or even thousands of customers creates candidate control groups with many more customers for each treatment to be compared. In addition, allowing such a huge ratio of control customers to treatment significantly improves the probability of getting a close match for each treatment customer.

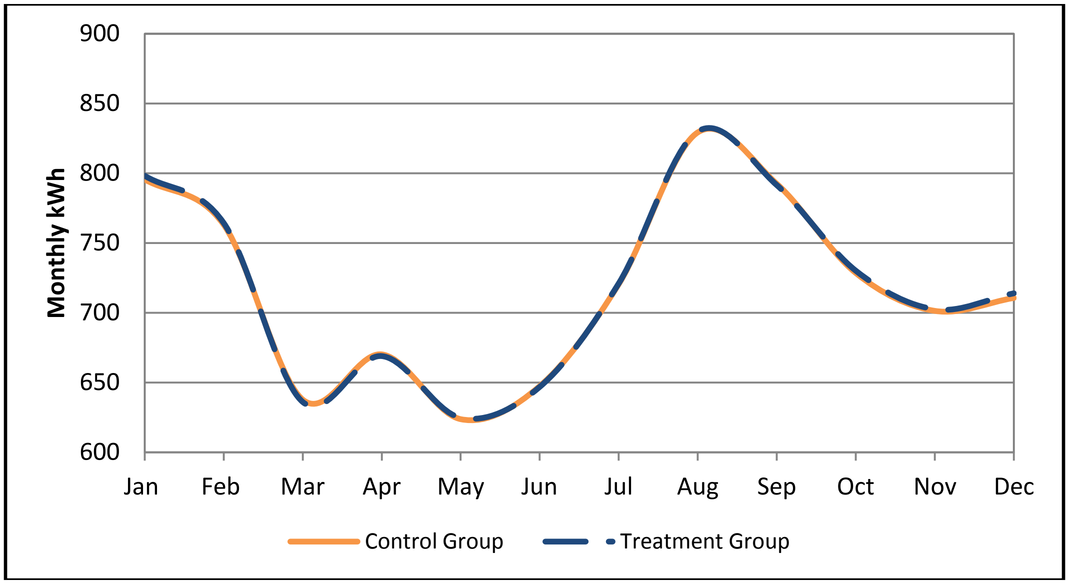

Figure 5: Energy Alerts Monthly Load Profiles - PY 2012Matched

After obtaining a satisfactory match a difference in differences method can be used to estimate program savings. This method will help to remove any pre-treatment differences in the two groups, as we have above in figure one, and isolate only the treatment period differences.

Strengths and Weakness of Each Method

Both the propensity score and the Euclidean distance methods have strengths and weaknesses, in some cases making up for the shortcomings of the other.

One of the great advantages to using a propensity score model is that they are computationally straightforward to estimate. Most statistical packages have procedures to estimate probit or logit models that can be used to create a propensity score for each customer. Because the score is created within a modeling framework, it is also simple to create many different matched control groups while experimentally employing different estimators in the model. Propensity score models also seem to work better in situations with smaller treatment and control group pools because while you can pre-stratify customers, a propensity model can also be estimated with the stratification variables in the model.

Propensity models also have some limitations, because the model is estimating the probability of participation, or the probability of a dependent value equal to 1, there must be a reasonable ratio of ones to zeros in the model in order for the regression to reliably estimate the propensity score. Depending on the potential control group pool, this may require taking a random sample of the control population to use in the modeling. This is limiting only insofar as it limits the control group pool, because the more candidate control customers there are to choose from the better the match.

The Euclidean distance metric also has its own strengths and weaknesses. One of the strengths of the method is that because it is not a model, there really is no limit to the ratio of treatment to control customers. The approach can be used with very large or very small groups. In our experience however, the method works best with larger groups and larger control pools because of the initial stratification. With smaller stratified groups, there are simply less control group customers to choose from which can affect the quality of the match. The method also allows the analyst greater control over the importance of the variables being used in the Euclidean distance through weighting. If for example, summer impacts are more important than winter impacts, weights can be applied to the summer usage variables to reflect that importance and a weighted Euclidean distance can be used. Conversely, when using a propensity method, the model determines the relative importance of the various inputs.

One of the biggest drawbacks of using a Euclidean distance method is the increased processing and coding required to execute the method. Unlike a probit or logit model the Euclidean distance does not have any pre-existing procedures within commonly used statistical packages and must be built from scratch. In addition, the datasets expand exponentially when comparing each treatment customer with every potential control group customer, therefore processing time of these very large datasets can be an issue. Even with powerful servers, it can take days of continuous processing to create the matched control groups when working with very large populations which can limit experimentation with different combinations of variables that might enable the best possible match. Table 2 below summarizes some of the pros and cons of each method.

Table 2: Pros and Cons of SPSM vs. SEDM

Stratified Propensity Score Matching		Stratified Euclidean Distance Matching	
Pros	**Cons**	**Pros**	**Cons**
Straightforward estimation in most statistical software	Must maintain a reasonable ratio of treatment to control customers	No limit to the ratio of treatment to control customers	Computationally intensive
Easily experiment with multiple model specifications	No control over relative importance of estimators	May work better with large treatment and control group pools	Must be coded from scratch not readily available technique
May work better for smaller treatment groups		Control over relative importance of estimators using weighting within Euclidean distance algorithm	

Recommendations and Other Considerations

Both the stratified propensity score matching (SPSM) and with stratified Euclidean distance matching (SEDM) methods for selecting quasi-experimental control groups can be implemented relatively easily and will provide good results under circumstances where a true randomized experimental design is not possible for the estimation of energy impacts of an energy conservation, energy efficiency, or demand response program. Whenever possible, however, randomized assignment of treatment and control groups should be utilized so as to best avoid the influence of selection bias.

There are a few recommendations that should be considered when using either the SPSM or SEDM methods. These are:

- The ratio of non-participants to participants should be at least 4:1
- When the non-participant population is extremely large (greater than 10 times the participant population), first randomly select a smaller non-participant population to use in the matching process.
- When the same control is the closest to two different treatment customers, calculate the second closest for all treatments and minimize the total overall distance between treatments and controls
- Use selection with replacement when the ratio of non-participants to participants is less than 4:1. It is necessary to weight the non-participants by the number of times they appear in the final analysis so as not to over inflate the standard errors of the estimates.

The situation can occur where the non-participant population is extremely large relative to the participant population. This can occur for any number of reasons, such as the program of interest is relatively new and not many customers have elected to participate as of yet. This is particularly an issue for SPSM as the logit model will not be able to estimate the probability of participation with such a large non-participant population. In order for the logit model to work, it will be necessary to randomly select a smaller number of non-participants for modeling purposes. It is recommended that the ratio of non-participants to participants be limited to no more than 10:1.

The number of available non-participants for control group selection and matching needs to be large enough so that a good match can be found for each participant. When participants are self-selecting, they tend to be different on average from those who are not participants. As a result, it will be necessary to use a larger pool of non-participants than there are participants. Both of Itron's and EnerNOC's experience suggests that the ratio of non-participants to participants should be at least 4:1.

Situations often occur, both with SPSM and with SEDM, where the same control customer is the closest to two different treatment customers. In these cases there are a couple of options. EnerNOC has used the following approach to minimize the overall distance when employing a Euclidean distance match. The distance is calculated between the two treatment customers and their second closest matches, and then the control customer is assigned to the treatment customer with the poorer alternate match (i.e. larger distance to their second closest match). This ensures that when a control is matched to two treatments, the treatment customer who is farthest from their second choice gets their first choice, minimizing the total overall distance between all treatments and controls. An example of the matching strategy is shown below in Figure 6 for two dimensions (average summer and winter weekday usage). This is done for illustrative purposes, but please note that this is equally valid in multiple dimensions.

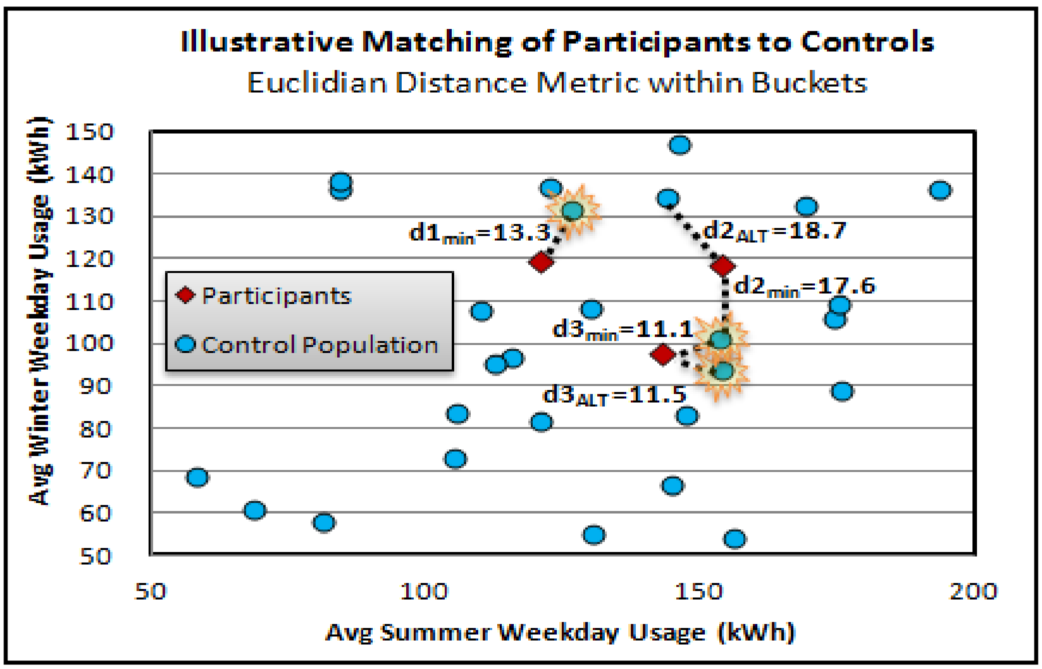

Figure 6: Illustrative Example of Matching Strategy

In the example above, Participant 1 finds its closest match in the control population to be a distance of 13.3 units away.[3] Participant 2 and 3, however, both have their minimum distance criteria satisfied by the same Control point. The next best alternative match for Participant 3 is 11.5 units away, whereas Participant 2's next best alternative match is 18.7 units away. Because Participant 3 can go a shorter distance to its alternative, it graciously steps aside to give Participant 2 its first choice of Control match, and it gets its second choice, 11.5 units away. Once every participant in the treatment group is matched with a unique customer from the control population, the comparison analysis can proceed.

[3] The units here happen to be kWh, but it is more appropriate and easier conceptually to think of them as generic units of distance.

Itron has developed a process when there are many more non-participants with the same propensity score than the participants. When this is the case, nonparticipants are randomly selected from those available to match with the participants. For example, if there were 10 participants with a propensity score of 0.200, and there were 15 nonparticipants with the same propensity score, 10 nonparticipants were randomly selected from the 15 similarly scored sites. This process was done first at the 0.001 decimal level. Matching at the thousandth digit was successful for the vast majority of participants. The process was iterated at the 0.01, 0.1, and 0.2 levels for any remaining unmatched participants.

In some situations, the ratio of non-participants to participants is less than 4:1 and approaching 1:1. In these instances it is necessary to make the best of the situation and use selection with replacement. What this means is that a non-participant may be selected more than once as a control. The non-participant is selected once as the closest fit and then they are put back into the pool to be available to be selected as a control for another participant.

In conclusion, when using either of these control group matching methods we recommend that a) the ratio of non-participants to participants should be at least 4:1, b) when the non-participant population is extremely large (greater than 10 times the participant population), first randomly selection a smaller non-participant population to use in the matching process, c) when the same control is the closest to two different treatment customers, calculate the second closest for all treatments and minimize the total overall distance between treatments and controls, and d) use selection with replacement when the ratio of non-participants to participants is less than 4:1.

References

Rosenbaum, Paul R.; Rubin, Donald B. (1983). "The central role of the propensity score in observational studies for causal effects". Biometrika 70 (1): 41–55. DOI:10.1093/biomet/70.1.41.

Elizabeth A. Stuart A. (2010). “Matching Methods for Causal Inference: A Review and a Look Forward”. Statistical Science, Vol. 25, No. 1, 1–21. DOI: 10.1214/09-STS313. © Institute of Mathematical Statistics.