Finding Your Perfect Match: A Qualitative Investigation of Matching Approaches

Abigail Nguyen, Kelly Marrin, and Katherine Chiccarelli, Applied Energy Group, Walnut Creek, CA
Andrew Lee, Pacific Gas & Electric, San Francisco, CA

ABSTRACT

Inspired by Gary King's scholarly papers on matching, we investigate the relevance and applicability of his findings to the energy industry as they relate to the two leading methods in quasi-experimental design: propensity score matching (PSM) and Euclidean distance matching (EDM). King postulates that propensity score approaches should not be used for matching in favor of Euclidean distance. We tested his theory using real utility data for a diverse, residential population by simulating EM&V analyses on three typical demand response and behavioral programs: a time-of-use rate, an AC cycling program, and a home energy report program. These three programs cover impact evaluations at both the hourly and monthly levels.

The main goal of our study was to determine if there are specific conditions under which one matching approach is significantly better at estimating program impacts. By using three different simulated programs we could test the effect of a variety of conditions on the quality of the match including: data frequency (interval vs. monthly), program type (behavioral, event driven, or pricing), and ratio of treatment to control customers. In each case, we evaluated the performance of the two matching methods for both accuracy and bias to determine if one technique yields a better match. Our findings support both methods: the time-of-use rate did not show preference to either method, but the AC cycling performed best with PSM variations, and the home energy report program showed promising results using an EDM approach.

Introduction

Simulated Programs

First, we discuss the three types of demand response and behavioral programs we simulated in this study. We chose these programs not only because they are typical in the industry, but also because evaluating these programs requires the application of common approaches typically encountered in impact evaluations.

Time-of-Use Rate. A Time-of-Use (TOU) rate plan is a program that prices energy use based on the changing demand throughout the day and season. Essentially, prices are highest during peak periods and lowest during off-peak periods. Participating customers are matched with control customers using a pre-treatment timeframe, where both groups are not affected by a time differentiated rate.

Air Conditioner Cycling Program. An air conditioner (AC) cycling program is an event-based program. Participating customers receive a device to be installed to their AC unit, which can be remotely controlled by their electric utility to run at a lower capacity during an event. In this type of program, the treatment period is solely comprised of event days, and so the pre-treatment period, that is used to match treatment to control customers, is comprised of a group of non-event days that are similar to event days in terms of weather, weekday and season.

Home Energy Report Program. A Home Energy Report (HER) program is a behavioral program that sends participating customers a report on their energy usage, typically every month. These reports aim to

increase customer awareness of their energy habits and encourage changes in their usage patterns. Similar to the TOU program, participating customers are matched with control customers using a pre-treatment timeframe, where both groups are not receiving an HER.

Analysis Methodology

Our simulations consisted of the following aspects:

- Establish the population based on available data required for each simulated program,
- Assign the population to the treatment group and control group pool,
- Simulate program impacts during a treatment period,
- Match treatment customers to control customers using both Euclidean distance matching (EDM) and propensity score matching (PSM),
- Estimate program impacts using a difference-in-differences (DID) approach
- Evaluate the quality of the match by comparing estimated impacts to "true" or simulated impacts and testing the matched control groups.

Establish the Population. For this study, we used control group pool data from previous analyses. It was important to use customers that did not participate in any programs to ensure that we had clean and unbiased data. For example, we did not want to include customers with large dips in their daily loads (like a DR program) or customers with substantial year-to-year differences in their seasonal loads (like a behavioral program).

Treatment Group and Control Group Pool Selection. For each simulated program, we started off with a population of customers, and first strategically selected treatment customers. A simple random sample, because of the nature of randomization, would create a treatment group and a control group pool that are statistically perfect for each other. This would result in well matched control groups regardless of the methodology. Thus, we selected a treatment group with the program characteristics in mind. For example, the simulated TOU treatment group was selected from the customers in the top 50 percentile of the population's summer usage.

The remaining population is the control group pool. We typically prefer at least 1:10 treatment to control pool ratio, which we were able to implement in the HER simulation. However, given the constraints of our available hourly usage data, we ended up with a 1:3 treatment to control pool ratio for the TOU and AC cycling simulations.

Simulate Program Impacts. To simulate impacts for each program, we looked at previous impact evaluations and obtained the hourly or monthly percentage impacts for each program. We distributed these impacts into low, medium, and high savings to introduce some variability into our data. We then randomly assigned our treatment group into three equal-sized groups: low, medium and high savers; and applied each set of percentage impacts to the corresponding group.

Matching Methods. Below, we describe the two matching methods we explore in this study: PSM and EDM. The two methods differ in the formulation of the distance metric, however, once the metric is established, the match selection process is the same. For each treatment customer, we calculate their distance to every control customer. The treatment and control pair with the smallest distance is considered a match. If a treatment customer shares their best match with other treatment customers, we look at their second "closest" match to determine who "wins" their first match. These pairs are removed from the pool, the process repeats until all treatment customers are matched with a control customer.

Euclidean Distance Matching. The Euclidean distance metric is the straight-line distance between two points in Euclidean space. The ED metric is a form of the Mahalanobis distance metric that King refers to in his several papers on matching. It is defined as the square root of the sum of the squared differences between the matching variables. Any number of relevant variables can be included in the ED metric. Equation 1, below, shows an example of an ED metric using twelve months of usage data.

$$ED = \sqrt{(jan_{Ti} - jan_{Ci})^2 + (feb_{Ti} - feb_{Ci})^2 + \ldots + (nov_{Ti} - nov_{Ci})^2 + (dec_{Ti} - dec_{Ci})^2} \quad (1)$$

Because of the strictly quantitative nature of the ED metric, it is difficult to incorporate qualitative variables within the distance metric. Using simple indicator values, such as 0 and 1, for qualitative variables alongside quantitative variables, such as energy usage, can inadvertently give weight to any of the variables. Although this can be resolved by placing weights within the ED metric, segmentation is another way to incorporate qualitative variables in a strictly quantitative metric like the ED metric. When segmentation is applied, we are now simulating a randomized block design (RBD) instead of a randomized control trial (RCT). In a RBD, the defined population is first divided into blocks and then each block is randomly assigned into treatment and control groups.

Propensity Score Matching. The PSM methodology is the most commonly used matching method in observational studies, and possibly in the energy industry. It develops a propensity score (PS) using a logistic regression model that attempts to estimate the probability of receiving treatment given a set of covariates or what we have referred to as matching variables. Equation 2 shows an example of a logistic regression model used for PSM that includes both quantitative and qualitative variables: twelve months of usage data, climate zone, and home dwelling type. From this model, we obtain the PS and use it as the metric in the match selection process.

$$Treated_i = \beta_0 + \beta_1 jan_i + \beta_2 feb_i + \cdots + \beta_{12} dec_i + \beta_{13} climate_i + \beta_{14} home\ type_i + \varepsilon_i \quad (2)$$

Where

$Treated_i$ is an indicator variable that takes on the value of 1 if customer i is a treatment customer or the value of 0 otherwise,

β_0 is the model intercept,

$\beta_1 jan_i + \beta_2 feb_i + \cdots + \beta_{12} dec_i$ are the twelve monthly usage values for customer i,

$\beta_{13} climate_i$ is the climate zone where customer i is located,

$\beta_{14} home\ type_i$ identifies whether customer i is a single-family or multi-family home, and

ε_i is the error of the model.

Variations Used in this Study. In each simulated program, we used three basic variations of the matching methodology: a segmented EDM, a segmented PSM, and a basic PSM (i.e. not segmented). We describe each scenario below.

- **Basic PSM** – customers are not segmented prior to matching, and the propensity score incorporates region (inland or coastal) and dwelling type (single- or multi-family) and usage data during pretreatment periods (i.e. average on- and off-peak hours on weekdays for TOU, average event and non-event windows on event-like days for A/C cycling, monthly usage for HER)
- **Segmented PSM** – customers are first segmented by region (inland or coastal) and dwelling type (single- or multi-family), and the propensity score incorporates only usage data during pretreatment periods (i.e. average on- and off-peak hours on weekdays for TOU, average event and non-event windows on event-like days for A/C cycling, monthly usage for HER)

- **Segmented EDM** – customers are first segmented by region (inland or coastal) and dwelling type (single- or multi-family) prior to matching, and the Euclidean distance metric incorporates only usage data during pretreatment periods (i.e. average on- and off-peak hours on weekdays for TOU, average event and non-event windows on event-like days for A/C cycling, monthly usage for HER)

For the TOU and AC Cycling simulations, we included another layer of testing by incorporating all 24 hours of data, and therefore have a total of six variations tested for the TOU and AC Cycling simulations.

Estimation of Impacts using a Statistical DID. After the matched control groups for each variation of the two methodologies are selected, the simulated impacts are estimated using the difference-in-differences (DID) method. The DID compares the hourly or monthly usage of the treatment customers to the matched control group customers, both during the participation period (treatment period) and for a time before participation started (pretreatment period). Comparison during the treatment period gives an unadjusted estimate of the impacts. This estimate is then corrected using the difference during the pretreatment period to adjust for any preexisting differences between the treatment and control groups.

Equation 3 shows a simplified form of the mathematical calculations used in the difference-in-differences analysis to estimate energy savings for each day type or month.

$$Savings = (Cntl_{after} - Tx_{after}) - (Cntl_{before} - Tx_{before}) \quad (3)$$

Where, $Cntl$ and Tx refer to the average control and average treatment group customers, respectively, and the subscripts $before$ and $after$ refer to the pretreatment and treatment periods, respectively

Comparison of Impacts and Testing the Matches. We performed the following tests on the matches. For all tests, we used the 10% significance level or α=0.10.

1. Compare the average pretreatment loads of the treatment and control groups:
   - Visual comparison of the loads shapes,
   - Two-sample t-tests to check for significant differences at an hourly or monthly level,
   - Mean absolute percent error (MAPE) to quantify the differences at the daily or annual level, and
   - Mean percent error (MPE) to check for directional bias at the daily or annual level.
2. Compare the estimated and simulated impacts and reference loads:
   - One-sample t-tests to check if the estimated impact/reference load is significantly different from the simulated impact/true reference load at the hourly or monthly level; in other words, we test if the simulated impacts are within the 90% confidence intervals determined by the estimated impacts, and
   - MAPE and MPE to quantify differences and check for directional bias at the daily or annual level. For the TOU and AC Cycling simulations, we also do this test for the on-peak and event windows, respectively.

Creating a Distribution of Savings. To test the strength of our conclusions, we performed each simulation 101 times to get a distribution of simulated and estimated savings. This process is called a Monte Carlo simulation. However, it is important to note that the quality of the savings estimate relies directly on the quality of the reference load, or baseline. So, we simulated 101 reference loads and analyzed their quality under each variation using the following tests:

- MAPE and MPE to quantify differences and check for directional bias, and
- One-way ANOVA tests to check for statistically significant differences between the variations implemented

Equation 4 below shows the mathematical calculation of the estimated reference for a statistical DID. The variables are as defined above in Equation 3.

$$Estimated\ Reference\ =\ Cntl_{after} - (Cntl_{before} - Tx_{before}) \quad (4)$$

RESULTS – TOU and AC Cycling Simulation[1]

Pre-Treatment Period. We first compared and tested average pretreatment period loads. Figure 1 shows both the visual comparisons and a visual representation of the two-sample t tests. The graph on the left shows the visual comparisons where the blue dotted line represents the treatment group's pretreatment period load shape, while the rest represent the matched control group for each corresponding matching methodology. The graph on the right shows the results of the two-sample t tests. In these tests, we are checking for statistically significant differences between the groups at each hour. A round marker indicates a significant difference (at the 10% level) between the treatment and indicated matched control group at the specified hour. Table 1 and Table 2 show the daily MAPE and MPE results.

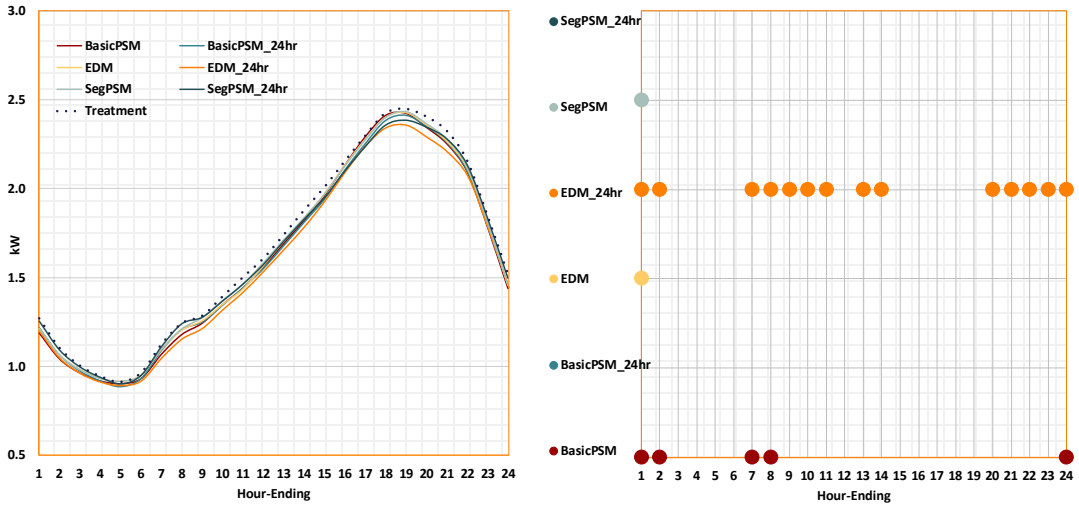

Figure 1. TOU Analysis: Pretreatment Match Comparison and Two-Sample t Tests

Visually, all approaches performed relatively well in both the TOU and AC Cycling simulations. We do see some poorly matched hours in the morning and the evening (EDM 24-hr, SegPSM 24-hr). Not surprisingly, these matching variations use the average hourly load shape and therefore give equal weight to each hour of the day. The variations that give more weight to the on-peak (4-9 PM) or event (4-7 PM) window (Basic PSM, EDM, SegPSM) show better-matching hours in those windows.

Based on the hourly t tests, the 24-hr basic and segmented PSM methods gave the best pretreatment matches for the TOU simulation, with matched control groups showing no statistically significant differences for all 24 hours. For AC Cycling, all variations except for average event and non-event window EDM showed statistically significant difference for Hour Ending 10 (HE-10). The average event and non-event window EDM variation showed no statistically significant differences for all 24 hours.

The variation with the best MAPE, which quantifies prediction accuracy, for both simulations is the 24-hr segmented PSM. An interesting observation is that the MPE are all positive (with one exception),

[1] To adhere with the 12-page limit, figures that may seem redundant were omitted from this version. These figures will be shown during the presentation and are also available in the full white paper version.

meaning that all matched control groups have a slightly lower average pretreatment period load shape. On top of that, the both basic PSM approaches (for TOU) and 24-hr EDM (for both simulations) have equal MAPE and MPE, meaning that these matched control groups are lower than the treatment group on all 24 hours.

Table 1. TOU Analysis: Pretreatment MAPE (%) and MPE (%)

Hour	EDM	Basic PSM	Segmented PSM	EDM (24-hr)	Basic PSM (24-hr)	Segmented PSM (24-hr)
MAPE	2.23	3.12	1.91	4.48	2.92	1.57
MPE	2.18	3.12	1.86	4.48	2.92	1.57

Table 2. AC Cycling Analysis: Pretreatment MAPE (%) and MPE (%)

Hour	EDM	Basic PSM	Segmented PSM	EDM (24-hr)	Basic PSM (24-hr)	Segmented PSM (24-hr)
MAPE	1.63	2.98	1.85	2.33	1.27	1.15
MPE	1.13	2.94	1.50	2.33	1.10	(0.34)

Impact and Reference Load. Next, we show the comparisons and test results for the estimated impacts compared to the simulated impacts. The subsequent figures are as described above except that the graphs on the right now represent one-sample t tests instead of two-sample t tests. Table 3 and Table 4 show the MAPE and MPE results. For the impacts, we show the on-peak MAPE and MPE, while we show the daily MAPE and MPE for reference loads.

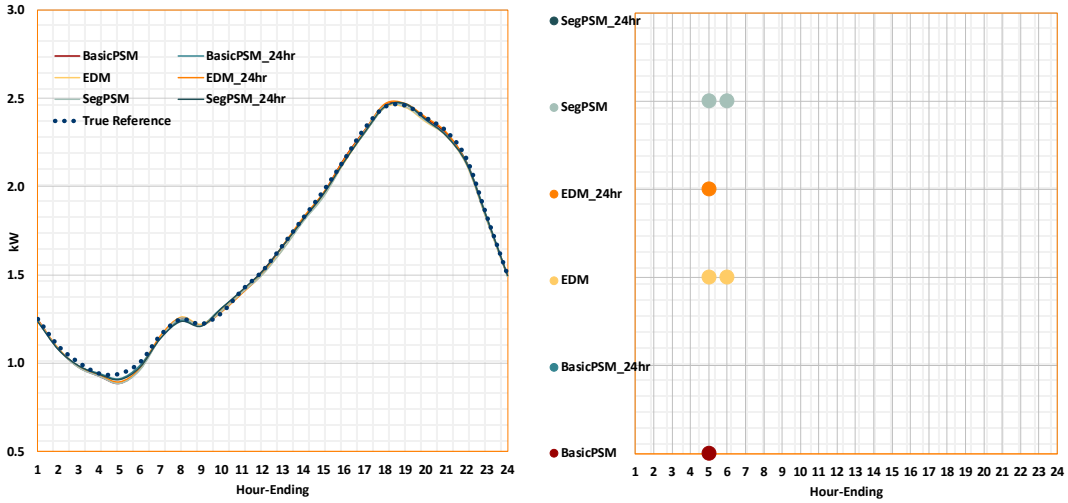

Figure 2. TOU Analysis: Reference Load Comparison and One-Sample t Tests

Table 3. TOU Analysis: Impact and Reference Load MAPE (%) and MPE (%)

Load	Statistic	EDM	Basic PSM	Segmented PSM	EDM (24-hr)	Basic PSM (24-hr)	Segmented PSM (24-hr)
Impact	On-peak MAPE	4.26	2.25	4.17	2.40	3.64	3.90
	On-peak MPE	4.26	0.98	3.78	0.15	2.98	2.91
Reference Load	Daily MAPE	1.36	1.06	1.46	1.04	0.89	1.05
	Daily MPE	1.19	0.90	1.34	0.87	0.69	0.84

First, we discuss the TOU simulation results. The visual comparison of estimated impacts appear to do well during the on-peak period and less so during the off-peak period. This is likely because the off-peak period has impacts very close to zero, which require larger sample sizes to improve estimates. Despite this, the off-peak period impacts still performed very well in the t tests, having only significant differences at HE-5 and HE-6. Because of the poor off-peak point estimates, the impacts MAPE and MPE are based only on the on-peak hours. Overall, the basic PSM approach gave the best MAPE, followed by the 24-hr EDM. As the pretreatment period tests showed, the impact estimates are also generally lower than the simulated impacts.

On the other hand, the poor accuracy shown in the off-peak impact estimation cannot be seen in the estimated reference loads. As shown in Figure 2, all six approaches estimated the reference load very well. The hourly t tests, MAPE, and MPE support this. Both 24-hr basic and segmented PSM approaches gave reference loads that contain the true reference load for all 24 hours at the 90% confidence level. All six approaches also gave excellent daily MAPEs, all under 1.5%. The standout approach, though, is the 24-hr basic PSM with a very low 0.89% MAPE.

Table 4. AC Cycling Analysis: Impact and Reference Load MAPE (%) and MPE (%)

Load	Statistic	EDM	Basic PSM	Segmented PSM	EDM (24hr)	Basic PSM (24hr)	Segmented PSM (24hr)
Impact	On-peak MAPE	4.28	2.65	5.03	4.88	1.79	4.46
	On-peak MPE	4.28	2.65	5.03	4.88	1.79	4.46
Reference Load	Daily MAPE	0.84	0.83	1.06	0.83	0.86	0.94
	Daily MPE	0.13	0.15	0.44	0.49	0.35	0.27

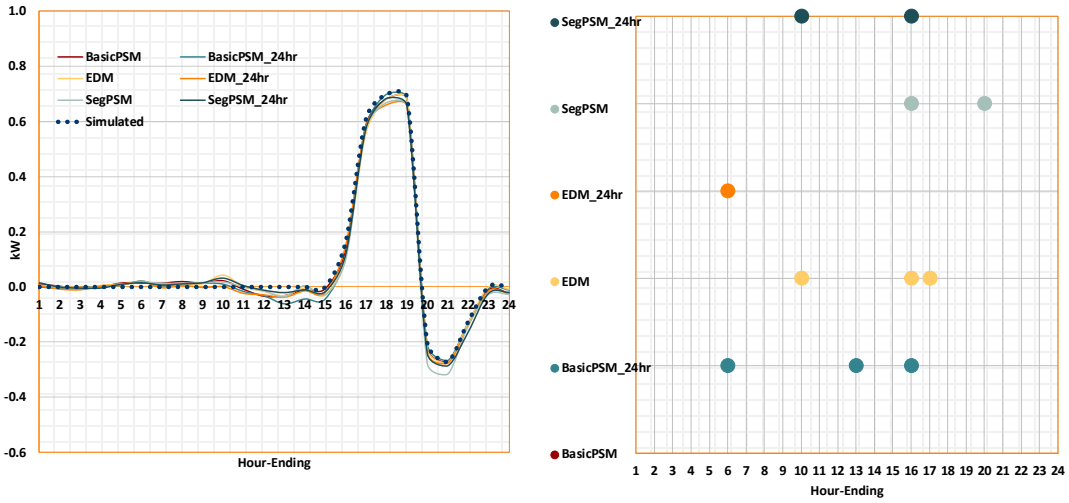

Figure 3. AC Cycling Analysis: Impact Comparison and Two-Sample t Tests

The AC Cycling simulation shows slightly different results compared to TOU. As shown in Figure 3, the estimated impacts appear to do well throughout the entire day. This is likely because the off-peak period has impacts very close to zero, which require larger sample sizes to improve estimates. However, the t tests do not show consistent results across the variations, showing differing sets of hours that did not contain the simulated impacts within their 90% confidence intervals. Following the TOU simulation, the impacts MAPE and MPE are based only on the event window hours. Overall, the 24-hr basic PSM approach gave the best MAPE, followed by the other basic PSM. Interestingly, all six approaches gave impact estimates that are lower for all 24 hours, shown by the MPE being equal to the MAPE.

On the other hand, the inconsistency cannot be seen in the estimated reference loads. Like the TOU simulation, all six approaches estimated the reference load very well. The hourly t tests, MAPE, and MPE support this. All six approaches gave reference loads that contain the true reference load for all 24 hours at the 90% confidence level. All six approaches also gave excellent daily MAPEs, with all except for one under 1%. For this simulation, there is no standout approach, with all MAPE results being very close together. The two with the smallest MAPE are the basic PSM and 24-hr EDM with 0.83% MAPE.

Monte Carlo Simulation Results. We performed each simulation 101 times to get a distribution of simulated and estimated savings, which we tested through estimated reference loads. The TOU results in Table 5 shows the average daily MAPE and MPE resulting from 101 estimated references loads for each approach, compared to 101 randomly selected treatment groups. Consistent with the results above, all six approaches gave excellent estimated reference loads, all having only slightly above 1% MAPE.

Table 5. TOU Analysis: Average Daily MAPE (%) and MPE (%)

Statistic	EDM	Basic PSM	Segmented PSM	EDM (24hr)	Basic PSM (24hr)	Segmented PSM (24hr)
MAPE	1.18	1.16	1.16	1.13	1.13	1.10
MPE	0.06	(0.09)	0.13	(0.04)	(0.03)	0.03

Unlike the initial results, the average daily MPE was not positive for all variations used. This is actually a very encouraging outcome. With half the MPEs positive and half negative and all very close to zero, we can be confident that all approaches do not exhibit any directional bias.

The one-way ANOVA test gave us p-values 0.51 and 0.47 for MAPE and MPE, respectively. Thus, we do not have enough evidence to show significant differences between the six variations of EDM and PSM that we explored.

Table 6. AC Cycling Analysis: Average Daily MAPE (%) and MPE (%)

Statistic	EDM	Basic PSM	Segmented PSM	EDM (24hr)	Basic PSM (24hr)	Segmented PSM (24hr)
MAPE	0.85	0.89	0.86	0.85	0.80	0.80
MPE	0.15	0.04	0.04	0.31	0.11	0.06

The Monte Carlo simulations for AC cycling gave slightly different overall results. Consistent with the results above, all six approaches gave excellent estimated reference loads, all only slightly above 0.8% MAPE (shown in Table 6). Unlike the TOU simulations, the MPE results are all positive, despite being very close to zero, suggesting a very small directional bias. However, considering the nature of this program, we selected the 10 hottest days as event days and the 10 closest in temperature as event-like days. Thus, the pretreatment (event-like day) correction between groups may not be a sufficient adjustment in programs such as this.

Also unlike the TOU simulations, the one-way ANOVA tests showed significant differences between the six approaches in both daily MAPE and MPE. Further testing on MAPE showed that both 24-hr basic and segmented PSM approaches were significantly different from the other four, suggesting that these two approaches have statistically better prediction accuracy. Also, the 24-hr EDM approach tested to have statistically different MPE compared to the other five approaches, suggesting that it had the strongest directional bias.

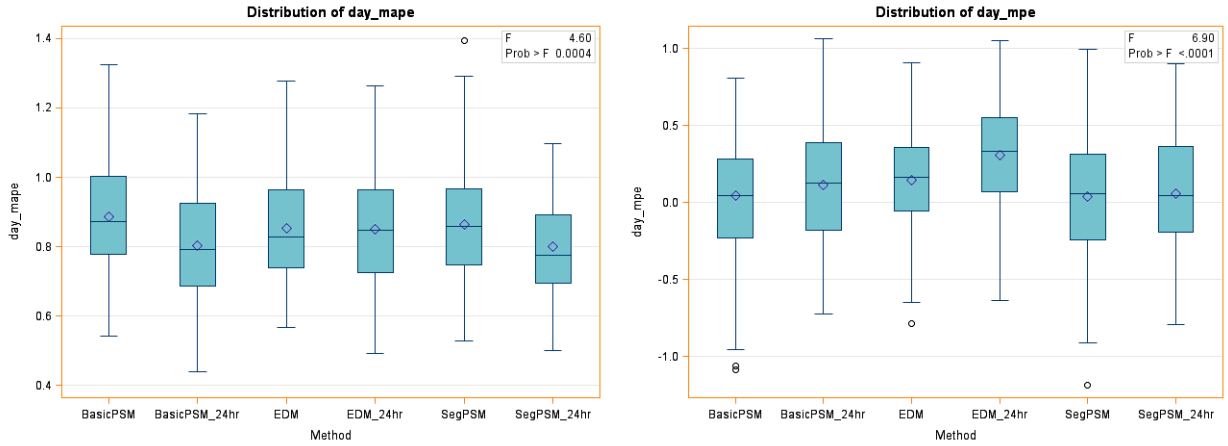

Figure 4. AC Cycling Analysis: One-way ANOVA on MAPE and MPE

RESULTS – HER Simulation

Before we discuss the HER simulation results, it is important to note a few key differences from the other two program simulations. First, we were not able to do a true basic PSM variation. Because of the amount of data involved in this simulation, the eligible control group pool had to be pared down for data processing reasons. Thus, the two PSM methods to be discussed going forward are both segmented PSM methods with the difference being that "Basic PSM" incorporates the qualitative variables (region and dwelling type) into the propensity score and the "Segmented PSM" does not. With that in mind, the "Basic PSM" results are likely better than that of a true basic PSM variation. This is something we'd like to explore further in future studies.

Also because of time constraints and the amount of data involved, we were not able to perform a Monte Carlo simulation for this program. Consequently, our results from this program simulation are not conclusive and more anecdotal. However, the Monte Carlo simulation is something we'd like to be able to complete as a follow up to our study.

Pre-Treatment Period. Figure 5 and Table 7 show the comparisons and test results done on the average pretreatment period loads. As with the results shown above, the graphs in Figure 5 show the visual comparison on the left and the visual representation of the two-sample t tests on the right. Both graphs essentially exhibit the same results: the EDM approach performed significantly better than the other two PSM approaches. Table 7 shows the annual MAPE and MPE results and affirm that the EDM approach performed the best among the three. It is also worth noting that both the EDM and segmented PSM matched control groups are consistently lower that the treatment group on average throughout the pretreatment period.

What's striking is that considering the size of the control group pool (higher treatment to control pool ratio), the expectation is that there is a better chance of a good match regardless of the chosen matching methodology. Also, these results are more in line with King's arguments regarding PSM methodology wherein he states that other matching methods, using Mahalanobis distance methods for illustration, are more successful overall.

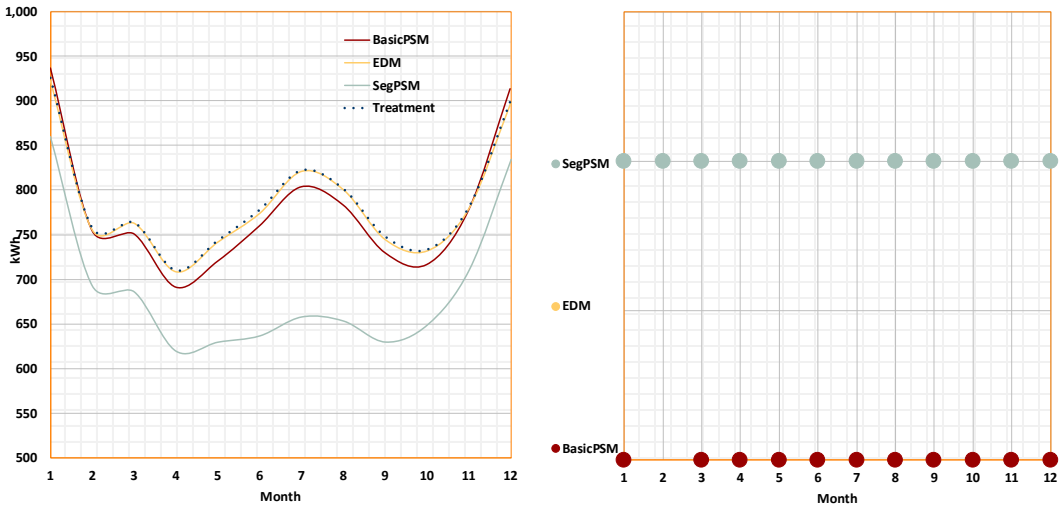

Figure 1. HER Analysis: Pretreatment Match Comparison and Two-Sample t Tests

Table 1. HER Analysis: Pretreatment MAPE (%) and MPE (%)

Hour	EDM	Basic PSM	Segmented PSM
MAPE	0.27	1.85	12.81
MPE	0.27	1.45	12.81

Impact and Reference Load. Next, we show the comparisons and test results for the estimated impacts compared to the simulated impacts. As with the results shown above, Figure 6 and Figure 7 show the visual comparison and one-sample t test results, while Table 9 show the annual MAPE and MPE for both impact and reference loads.

At first glance, the impact estimation results are very disappointing. The EDM approach still gives the most accurate month-to-month estimations with the best MAPE and MPE results, but overall, all three approaches performed poorly in estimating the simulated impact at the 90% confidence (one-sample t test results). What's promising, however, is that at the annual level, the EDM approach gave an estimate that is very close to the simulated impacts. Table 8 shows the annual simulated impact compared against the annual estimated impacts from each matching methodology (the sum of all statistically significant monthly point estimates).

Table 2. HER Analysis: Simulated v. Estimated Annual Impacts (kWh)

Simulated	EDM	Basic PSM	Segmented PSM
86.55	88.27	79.29	130.21

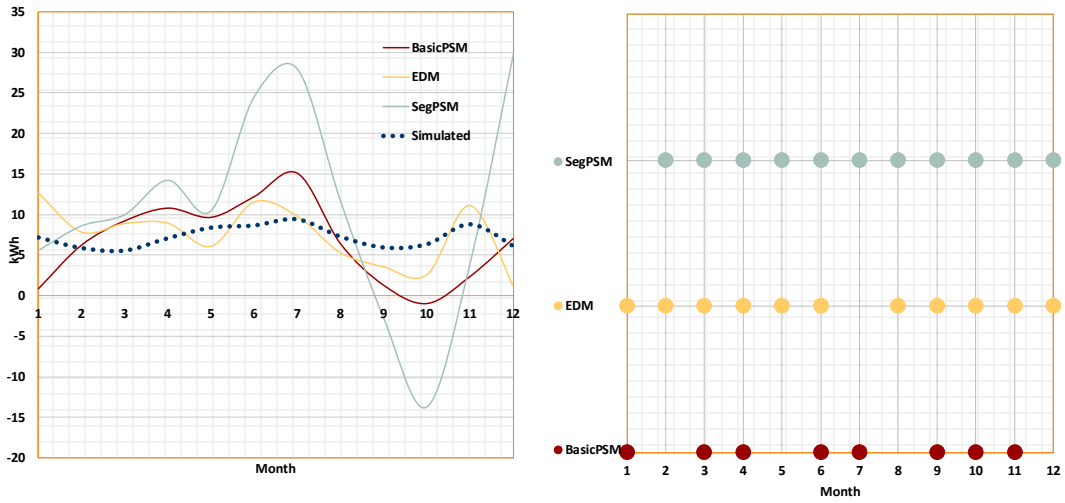

Figure 2. HER Analysis: Impact Comparison and One-Sample t Tests

The comparison and tests on the reference loads are more promising. The EDM approach is still the standout approach, but the visual comparison on the left of Figure 7 make the matched control groups look much better, even the segmented PSM group. However, both PSM approaches still performed poorly in the one-sample t tests (as shown on the right of Figure 7).

The annual MAPE and MPE results (in Table 9) for all three approaches do show excellent estimation accuracy, with all under 2% MAPE, and no indication of directional bias, with positive and negative MPE all close to zero.

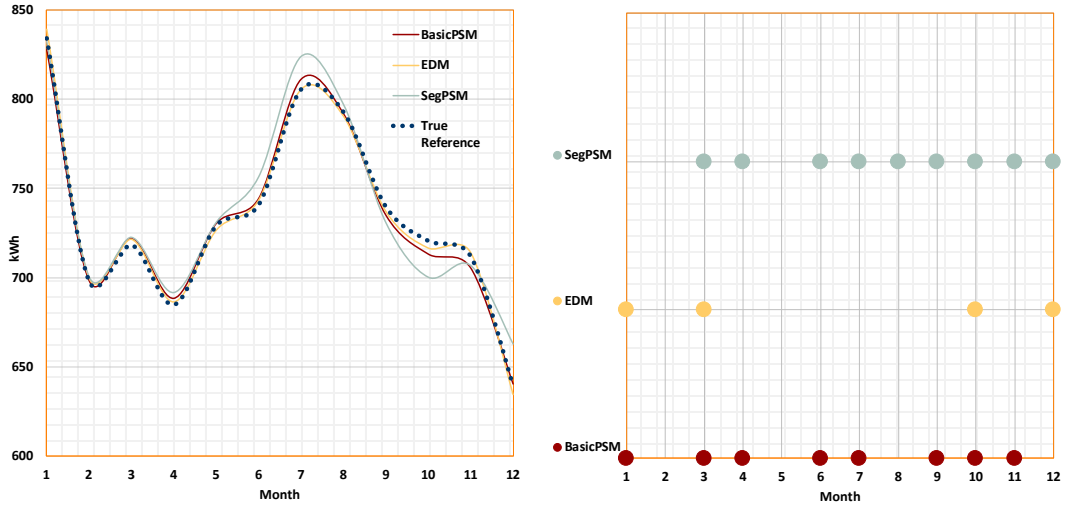

Figure 3. HER Analysis: Reference Load Comparison and One-Sample t Tests

Table 3. HER Analysis: Impact and Reference Load MAPE (%) and MPE (%)

Load	Statistic	EDM	Basic PSM	Segmented PSM
Impact	MAPE	41.27	52.15	135.05
	MPE	(2.14)	8.80	(44.93)
Reference Load	MAPE	0.39	0.50	1.32
	MPE	(0.02)	0.06	(0.52)

Key Findings

We did not find any significant differences between the six matching methodology variations for the TOU simulation. For AC Cycling, Basic and Segmented PSM using average event day hourly load (24-hr) showed significantly better accuracy than the other four. Also, EDM using average event day hourly load (24-hr) showed significantly worse directional bias.

Our initial conclusion is that any of the six variations are acceptable and will give sound results for similar impact evaluations using hourly usage data. All six variations gave excellent results in the pretreatment period, impact estimation, and reference load estimation. However, after seeing the HER results, we have reason to believe that our two hourly usage populations are inherently cohesive and will give excellent matching results regardless of the methodology. Since we used control group pool data from previous analyses, the availability of hourly usage data is a result of an initial match (segmented EDM) using monthly usage data.

The HER simulation, on the other hand, showed results that closely resembled King's claims. The Segmented EDM that simulated a RBD showed the most promising results overall. The pretreatment period comparison was almost spot-on, which is usually our only indicator in an actual impact evaluation. Reference load estimation was also impressive despite a few misses at the 10% significance level. Impact estimation was not as great on the monthly level, but still excellent at the annual level, which is often the bottom line in an impact evaluation. However, due to time and data processing constraints, we were only able to do this simulation once, giving us anecdotal results. A Monte Carlo simulation would give more statistically conclusive results.

Lessons Learned and Further Research

Because our HER simulation showed promising but not statistically conclusive results, we'd like to continue to look into this section of our study. One of the key constraints to a Monte Carlo simulation is the data processing constraint. We currently have 80,000 customers in our simulated treatment group and it's proving to be a sizeable group to work with for one simulation, so much more for multiple simulations. We set the treatment group size to be comparable to a typical HER program, but perhaps a power analysis would be appropriate to determine the most appropriate and efficient treatment group size for our population.

We'd also like to investigate our suspicion on the inherent cohesiveness of the hourly usage populations. Advanced Metering Infrastructure (AMI) systems are becoming increasingly available in the industry, and in turn, hourly usage analyses are becoming increasingly relevant. We see great value in having more conclusive results in this context. Lastly, we would like to explore how well each matching methodology performs at different subgroup levels of interest.

References

King, Gary and Richard Nielsen. Working Paper. "Why Propensity Scores Should Not Be Used for Matching". Copy at http://j.mp/2ovYGsW