Model Based Matching and Other Benefits of High Frequency Interval Data

Lullit Getachew, Jon Farland, Ken Agnew, Valerie Richardson, Gomathi Sadhasivan, DNV GL
Peter Franzese, California Public Utilities Commission

ABSTRACT

We explore two possible advantages of Advanced Meter Infrastructure (AMI) data for the development of matched comparison groups. First, we compare matching based on calendarized billing data versus true monthly data using AMI. Second, we investigate an alternative model-based matching, where parameter estimates from site-level information are used to generate matches, using both calendarized billing and daily AMI data. We hypothesize that energy consumption model parameter estimates that characterize the consumption dynamics of households may result in better matches than consumption-based matching. Such models that are based on daily AMI data, which affords improved visibility into consumption behavior of energy customers, may provide an even richer characterization of site-level usage that could be helpful when generating a program counterfactual. We find that matches based on calendarized billing data are practically as good as matches based on AMI data. Further, comparing balance using data from the same period where matching is done ('in-sample matching comparison') shows that model-based matching is of limited use. We also explore ways to determine the success of resulting matchings using nonparametric hypothesis testing, plots of matched samples, and propensity score diagnostics. The goal of this effort is to provide additional empirical and technical guidance when employing data for program evaluation, particularly when there is a need to set up program counterfactuals under less than ideal experimental designs.

Introduction

In this paper, we examine the effect data granularity and matching methodology have on matched comparison groups used to study the effect of participation in an energy efficiency program. We use data from residential Universal Audit Tool (UAT) participants of San Diego Gas & Electric (SDG&E) for this purpose. Branded as My Energy Survey, the tool provides customers with advice on energy efficiency, insight into areas of high energy use, and tips and suggestions for saving both energy and money based on responses to an online survey regarding household appliances, occupancy, and other dwelling characteristics. The purpose of the tool is to increase customer engagement and energy savings.

The first goal of our study is to understand the advantages Advanced Meter Infrastructure (AMI) data offer in matching, in general, and in differences in matching based on calendarized billing data versus true monthly data using AMI. A secondary goal is to understand the effect of alternative approaches in matching.

In energy efficiency and related studies, it is not uncommon for practitioners to conduct 5:1 matching using billing data to request AMI data to then develop a better 1:1 match. We use data from SDG&E's residential Universal Audit Tool opt-in program to explore if this a necessary undertaking. We use calendarized billing data from both participants and the non-participant residential population to generate an initial 5:1 matched dataset. We then use AMI data from this initial 5:1 matched population to generate best 1:1 matches. We repeat this exercise on the calendarized billing data to compare if the matches that AMI data yield differ from those based on billing data.

Outcome of the matching exercise, or check of balance following matching, is routinely based on hypothesis tests that employ test statistics to determine if matched comparison groups have statistically identical distributions. Such approaches are widely used and reported in studies where matched

comparison groups are used to study the effect of treatment intervention of all kinds. A group of scholars who specialize in matching, however, are critical of this approach and offer alternative and more appropriate means of testing balance. Our third goal in this paper is, thus, to examine what role tests of balance play in determining matching outcomes.

Matching Rationale

When studying the effect of treatment intervention, such as programs implemented to improve energy efficiency, the ideal setting is a randomized experimental design (RCT) where subjects are randomly assigned to treatment and control groups. Under such a setting the only differentiating factor between the two groups is treatment. Thus, difference in outcome pre- and post-treatment between these groups can be attributed to treatment.

In observational studies, where subjects self-select into treatment, treatment assignment is not random and may be tied to intrinsic characteristics of the subjects in this group. Estimated treatment outcome will, thus, reflect self-selection bias. In such cases, matching is a method that allows one to select opt-in (treatment) and comparison groups that are balanced in key characteristics. This method does not eliminate self-selection bias entirely, but it appears to be the best we can do to estimate the effect of treatment under quasi-experimental conditions.

In matched-comparison observational studies, we identify the effect of treatment assignment, $T = 1$, by evaluating average treatment effect on the opt-in (ATT) as follows:

$$ATT = E(Y_1|T = 1) - E(Y_0|T = 1) \quad (1)$$

Here, $E(Y_1|T = 1)$ is the expected outcome of treatment for the opt-in and $E(Y_0|T = 1)$ is the expected outcome of no treatment for opt-in individuals. The second term, however, is unobservable. If the expected outcome of comparison individuals is used in place of the second term, the average treatment effect on the opt-in (μATT) becomes:

$$\mu ATT = E(Y_1|T = 1) - E(Y_0|T = 0) \quad (2)$$

The difference between $\mu\boldsymbol{ATT}$ and ATT captures part of the selection bias. For example, those who self-select into using the UAT may already be motivated to save energy even in the absence of the tool for various reasons. Hence, the estimated treatment effect in this case reflects savings that occur, in part, because of such reasons. To the extent these reasons or characteristics are observable and non-time varying, matching based on them can provide us a counterfactual that reflects reduced self-selection bias. However, it will not solve the problem of self-selection bias entirely and estimates of treatment effect may still reflect self-selection bias of varying degrees.

Study Background

We employ two different data types (true monthly data based on AMI and calendarized billing data) and matching approaches that use propensity-score matching (PSM) to construct comparison groups. Such comparison groups generate counterfactuals that are used to evaluate the effect of engagement with the UAT. We are particularly interested in examining if AMI data facilitates a better matching approach than calendarized billing data.

AMI data collected at the hourly or sub-hourly level provides a certain degree of improved visibility into time-of-use behavior for both residential and non-residential sectors. While AMI data of energy consumption are smooth in aggregate, observations of electricity usage from single meters can be individually highly variable. The increased variation in AMI data can be well explained with appropriate site-level statistical models that provide a rich characterization of individual usage. Such a characterization affords a deeper level of understanding of customer-specific behavior and can be helpful when generating

a program counterfactual. This is especially true in situations where a RCT experimental design is not available.

This paper uses AMI data from about 10,000 SDG&E residential customers to explore matching outcomes that are used in analyzing the UAT program outcome. We use two different techniques for this purpose. The first matching technique uses pre-evaluation period consumption in a logistic model to generate propensity-scores. The second technique uses model coefficients from site-level consumption (PRISM) models in a logistic regression to generate another set of propensity-scores. The propensity-scores from both models are used to select matches for UAT participants using the nearest neighbor algorithm. We call the first method consumption-based matching and the second model-based matching.

We also apply these matching techniques on calendarized billing data of the same customers. Our primary goal is to examine whether higher frequency data leads to an improvement in matching. A secondary and related goal is to identify if the techniques used play a role in the matching outcome. Thus, to answer the research questions of the role AMI data and matching techniques play in matching we compare matching outcomes from a 2-by-2 set-up summarized in Table 1.

Table 1. Matching study matrix

	Data granularity	
	Monthly	AMI
Pre-period consumption	A	B
Modeled information	C	D

As the matching study matrix indicates, we compare matches from:

- The consumption-based method using monthly kWh (A)
- The consumption-based method using AMI data aggregated to monthly kWh (B)
- The model-based method using monthly billing data (C)
- The model-based method using AMI data (D)

Data

Consumption Data

We used monthly billing and AMI data from dual fuel and electric-only SDG&E residential customers that opted into the UAT in 2014 and had sufficient data in the matching. We defined data sufficiency for the analysis as participants that had 12 months of pre-evaluation period (the year 2012),[1] and 12 months of each pre- and post-participation period (the years 2013 to 2015) data. Since billing data reflect customer utility bills that do not align with calendar months we use a calendarized version of this data in the matching exercise. We generate weighted averages of monthly consumption using data from either side of billing month to allocate energy consumption to calendar months. Calendarized data allow us to compare consumption across matched groups that fall within the same month (time-frame).

We requested AMI data from SDG&E for five of the best matches for each participant identified by the consumption-based approach we outline in the Matching Methods Section below. We received 60-minute interval data in response. After examining the sufficiency of the data (the number of households

[1] We estimate impact models with errors clustered at the individual level routinely to address within-subject correlations (errors that are not iid at the individual level or monthly consumption values that are not independent at the household level). If we match using pre-treatment period consumption, we introduce error correlations between opt-in and matched-control subjects that we also would have to deal with when estimating standard errors. Thus, we match on consumption data outside of the evaluation period or on consumption data a year prior to the pre-treatment period to avoid this.

for which non-missing AMI data is available in the matching year of 2012), we determined we could apply matching using AMI data for dual fuel and electric-only customers of SDG&E. We present the number of households used in matching for each data type and method in Table 2. The number of control pool candidates available for matching are generally 5 times the matched-control numbers presented in the table. The matched-control numbers differ across the method due to data screening.

Table 2: SDG&E electricity participant numbers used in matching

	model-based matched household numbers		consumption-based matched household numbers	
	AMI data	Billing data	AMI data	Billing data
SDG&E dual	7,268	7,274	7,143	7,275
SDG&E electric-only	2,666	2,671	2,528	2,672

Customer Information Data

We stratified matching by climate zone using general customer information available from SDG&E. The general customer information database provides one of the 16 climate zones in which each of SDG&E's electricity customers are located. These zones are defined by the California Energy Commission (CEC) to reflect regional weather conditions. We consolidated CEC's classifications into three climate zones indicating desert, inland, and mild climate conditions. Table 3 presents where CEC's classifications fall in the three groupings and the number of participants within each climate zone used in matching.

Table 3: Climate zone groups for stratified matching

Climate zone group	Title 24 climate zone	SDG&E participant counts
Desert	15	12
Inland	8, 9, 10, 11, 12, 13, 14	3,634
Mild/Coastal	1, 2, 3, 4, 5, 6, 7, 16	6,343

Weather Data

We also used weather data (in the form of degree days) for the model-based matching approach, which we describe further in the Matching Methods. NOAA weather data were matched to premises based on Euclidean distance matching by zip code. For each weather station, we matched hourly dry-bulb temperatures with site-level interval consumption data.

Matching Methods

We used two different techniques to produce matched comparison groups using the billing or monthly consumption and AMI data. As Ho et al. (2007, 2011), Stuart and Rubin (2007), and Stuart (2010) illustrate, causal models and inference that use matched data need fewer assumptions about model specification and, thus, result in model estimates that are less prone to specification error. In the absence of a RCT, they outline matching approaches that can be used in quasi-experimental or observation studies to preserve the specification error reduction advantages of RCT. The matching approaches they provide

are based on propensity score matching (PSM), which we follow in our work here. The PSM process we follow based on their work involves the following general steps that we used in this evaluation:

1. Select subjects' characteristics that are related to treatment assignment.
2. Examine the distributions of these characteristics and exclude observations of the comparison group where these do not overlap as a first round of identifying common support for matching.
3. Fit a logistic regression model using these variables to estimate the probability that each subject gets assigned to the treatment group.
4. Conduct a second round of trimming or common support identification based on propensity scores.
5. Select a matching method, the number of controls in the many-to-one matching, and whether to match with or without replacement; match opt-in subjects' scores to non-treated (comparison) subjects based on these selections.
6. Conduct diagnostic checks to see selected matches are well-balanced.

The first step in the general PSM framework we present above involves the selection of characteristics that affect treatment assignment. The first matching technique uses pre-evaluation period consumption while the second uses site-level model parameters. We call the first consumption-based matching and the second model-based matching.

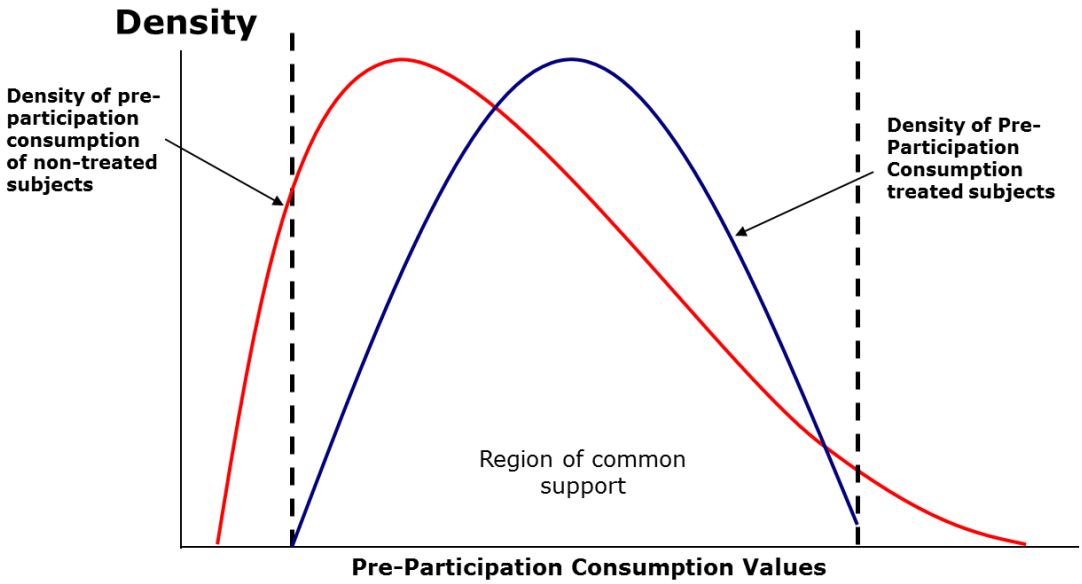
Consumption-Based Matching

To avoid the correlation of errors between treatment observations and their matches, we match on variables other than the dependent variable that includes both pre- and post-participation period consumption. Such variables can include any characteristics such as household size, heating and cooling source, and rate groups that may affect treatment assignment. They can also include variables measured before participation, such as pre-evaluation period consumption data.[2] For the consumption-based technique, we took the latter approach as comprehensive data on household characteristics were not readily available. Specifically, we used monthly data from the year 2012 that pre-date any consumption data that were used in the savings models. We also used climate zone information to stratify the data for matching. This involved implementing the matching procedure within three pre-defined climate zones for California - mild (coastal), inland, and desert.

Prior to estimating a propensity score model, we identified a first round of common support for matching by trimming the data based on the distribution of pre-participation consumption. Variable values of the comparison subjects that do not overlap with the values of the opt-in subjects were trimmed. In all the cases where we undertook matching, trimming pre-participation consumption values of the comparison subjects that are outside of the 1st and 99th percentiles resulted in the overlap of the distribution of these values with those of the opt-in. Figure 1 provides an example of how we established a region of common support.

[2] As we note in footnote 1, we use consumption data prior to the pre-treatment period for matching to avoid error correlations between matched pairs when estimating treatment impact.

Figure 1: Region of common support for matching



We fitted a logistic regression model using data that reflects common support and used the propensity scores from the regression to find matches for each opt-in subject based on k: 1 matches. The model is given by:

$$ln\left(\frac{(p)}{(1-p)}\right) = \beta_0 + \sum_t \beta_t X_t + \varepsilon \qquad (3)$$

Here, $p = p(T = 1|X)$ is the probability of receiving treatment (participation) and X is a characteristic variable or set of variables including pre-participation monthly consumption and consumption model parameters. The estimated propensity scores from this model were then used to establish a second-round of common support by trimming values of the comparison group whose scores are above the maximum and below the minimum of those of the opt-in subjects.

We used the nearest neighbor matching (NN) algorithm for this purpose. The approach produces matches for each opt-in subject, selected in random order, by searching for k propensity scores from the comparison group that are nearest to those of the opt-in subject's. We selected matches without replacement. Thus, a comparison subject selected as a match for a given opt-in subject was not available for matching again. This sort of matching is called 'greedy' because matches are made by only looking at distances of scores of randomly selected opt-in versus comparison subjects. Optimal matching, on the other hand, considers the overall distance between opt-in and comparison scores to select matches. The matches generated using either, however, are equally well-balanced.

Initially, we selected 5 best matches ($k = 5$) to identify an oversized matched comparison group for which to request AMI data. Within the 5:1 matched comparison group, we identified the optimal 1:1 matched comparison group for final models. As with the 5:1 matched comparison group selection, the 1:1 matched group was selected by identifying a comparison subject whose propensity score is closest to that of an opt-in subject selected randomly. Once selected, a comparison subject was not available for matching with any other opt-in household. We also conducted a matching exercise using AMI data for five of the selected comparison subjects that we discuss next in the next section.

Model-Based Matching

For model-based matching, we used a simple daily energy consumption model that estimates a set of regression models of energy use as a function of weather for each site (household) in the study[3]. The regression equation for a given site is given by:

$$E_t = \beta_0 + \beta_h H(\tau_h) + \beta_c C(\tau_c) + \varepsilon_t \quad (4)$$

where E_t = Energy, measured in kWh, used at time-period t.
$H_t(\tau_h)$ = Calculated heating degree days using actual observed temperature at time-period t and its deviation from reference temperature, τ_h.
$C_t(\tau_c) =$ Calculated cooling degree days using actual observed temperature at time-period t and its deviation from reference temperature, τ_c.
$\beta_0, \beta_h, \beta_c =$ Regression coefficients measuring the marginal effect of base load, heating load, and cooling load, on a single site's energy consumption, respectively.
$\varepsilon_t =$ Regression residual in time-period t.

A PRISM analysis uses cooling and heating degree-days to measure the variation in a site's energy consumption that can be attributed to variation in weather conditions. These cooling and heating variable constructs are calculated using the following equations:

$$C(\tau_c) = \begin{cases} 0, & x_t - \tau_c < 0 \\ x_t - \tau_c, & x_t - \tau_c \geq 0 \end{cases} \quad (5)$$

$$H(\tau_h) = \begin{cases} \tau_h - x_t, & x_t - \tau_h < 0 \\ 0, & x_t - \tau_h \geq 0 \end{cases} \quad (6)$$

In other words, if the observed temperature is above the cooling threshold τ_c, then that difference in degrees Fahrenheit is calculated as cooling degree days and vice versa for heating.

If the consumption data is utility billing data, the heating and cooling degree days for a billing period is traditionally given by calculating the heating or cooling degree days for each day within the billing period and aggregating across all days. The aggregation of degree days is then associated with time t.

We applied the PSM procedure to household-level model coefficients generated using the PRISM model given in (4) along with an estimate of the model's goodness-of-it (such as adjusted R-square). The steps involved in the model-based matching were:

1. Estimate energy consumption as a function of optimal heating- and cooling-degree days (HDD and CDD, respectively) using billing or AMI data.
2. Obtain base load estimates; HDD and CDD effects; estimates of optimal HDD and CDD bases; and model goodness-of-fit for each household.
3. Apply the PSM procedure outlined above using model coefficients to obtain 1:1 matches out of the 5:1 preliminary match comparison group for which we requested AMI data. The two different approaches that result in 1:1 matches are each optimally matched comparison groups from the same set of 5:1 households. To avoid correlation between treatment selection and outcome, by construction, we match on variables other than the dependent variable

[3] Like the PRISM model discussed in "The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures", Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol, 2013, National Renewable Energy Laboratory (NREL), https://energy.gov/sites/prod/files/2013/11/f5/53827-8.pdf

Test of Balance

The final step in the matching process was to check that generated matches are well-balanced. Checking matches are well-balanced involves ascertaining the distributions of the variables of the comparison and opt-in subjects (on which matching is done) are the same.

Hypothesis test of balance using methods such as t-test of difference in sample means and Kolmogorov–Smirnov (KS) test of equality of distributions of two-samples is routine in many empirical studies of matching. Ho et al. (2007) and Imai et al. (2008), however, indicate that such tests are inapplicable in the context of matching for several reasons.

First, matching is not an estimation that reflects the characteristics of some hypothetical population, but is strictly a feature of a given sample. Balance after matching is a property of the sample and not that of a broader population. Hypothesis tests of balance imply we can make inference about a population using sample data. Second, and relatedly, matching does not have a threshold below which the level of imbalance is acceptable. All undertakings that improve matches, up to producing perfectly matched samples, are useful. Hypothesis tests, on the other hand, imply that matches that meet a certain threshold are acceptable. Third, test statistics used in hypothesis tests of matching, including the t-test statistic,[4] are both a function of the condition of balance and the number of observations used to compute them. The power of the test improves as increasing number of control variables used in matching are dropped indicating improvement in balance even if such improvement does not exist.

Instead, they suggest tests of balance by directly comparing the empirical distribution of covariates used in matching via plots of their distribution and propensity score diagnostics. Such diagnostics include evaluating the mean difference in propensity scores using the standardized difference proposed in (Austin & Mamdani, 2006). For continuous variables, the standardized difference is given by $d = \left(\bar{X}_{treatment} - \bar{X}_{comparison}\right) / \sqrt{\left(S^2_{treatment} + S^2_{comparison}\right)/2}$ and is independent of sample size. A standardized difference that exceeds the value of 0.2 shows imbalance (Sawilowsky (2009)). However, the lower the value of standardized difference, the better the balance. Another diagnostic check of balance includes the examination of the ratio of the variances of the propensity scores in the two groups. A value that is close to 1 indicates balance whereas values that are close to 1/2 or 2 indicate extreme imbalance.

Based on the above, we determined balance by examining the distribution of unmatched and matched pre-participation consumption for opt-in and comparison subjects using histograms. In addition, we examined the quality of matches using both a hypothesis test of balance and the alternative diagnostics checks suggested by Ho et al. (2007) and Imai et al. (2008). We are interested in the outcomes from each set of approaches and in examining the applicability of the hypothesis tests of balance.

For the first approach, we tested the quality of the matches using the Kolmogorov–Smirnov test (KS test), which is a nonparametric test that examines the equality of the (cumulative) distributions of two samples. Under the null hypothesis of equality between the distributions, the KS test allows us to determine if the matched samples are statistically identical or not. For the second, we examined the standardized differences and the ratios of the variances of the propensity scores in the opt-in and comparison groups from the unmatched and matched datasets.

Results

As we discuss in the Matching Methods Section, we selected 1:1 comparison-to-opt-in households in our matching process out of the initial outsized 5:1 matches generated to request AMI data. In other

[4] This test statistic is given by $(\bar{X}_t - \bar{X}_c)/\sqrt{(S_t^2/N_t) + (S_c^2/N_c)}$, where the subscripts t and c signify treatment and comparison groups, $\bar{X}$, S^2 and N indicate the average value, variance and number of observations of the covariate or propensity score from the matching.

words, we selected the best match for each opt-in household in our study frame of the initial 5:1 matched universe. The best matches were selected using the matching-techniques we presented in the Matching Methods Section based on calendarized billing and AMI data. We ascertained balance in our resulting matches using three approaches.

First, we examined the distribution of the matched data for both dual-fuel and electric-only SDG&E households. We present the results for the consumption- and model-based billing data matches for dual-fuel households in Figure 2, and for consumption- and model-based AMI data matches for dual-fuel households in Figure 3.[5] Visual inspection of the figures makes it evident that the samples are well-balanced (matched). The values of consumption for the opt-in and matched comparison groups are very close across the entire consumption range in each case. Although the figures indicate balance, they do not allow us to rank the outcomes of the matches from the various matching approaches we used.

Figure 2: Consumption- and Model-Based Billing Data Matched kWh for Dual-Fuel Households

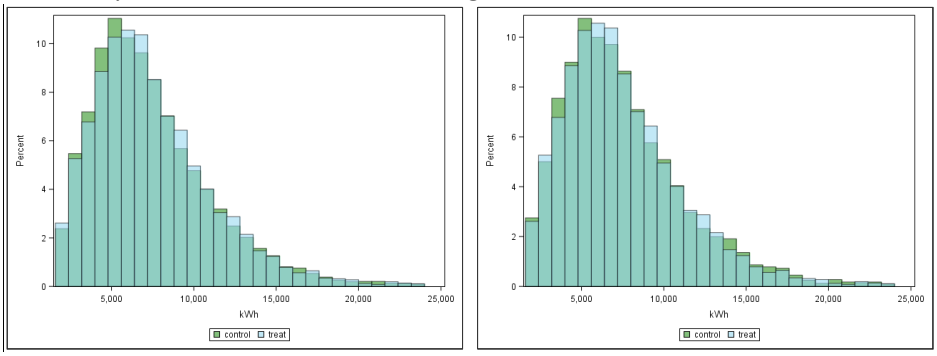
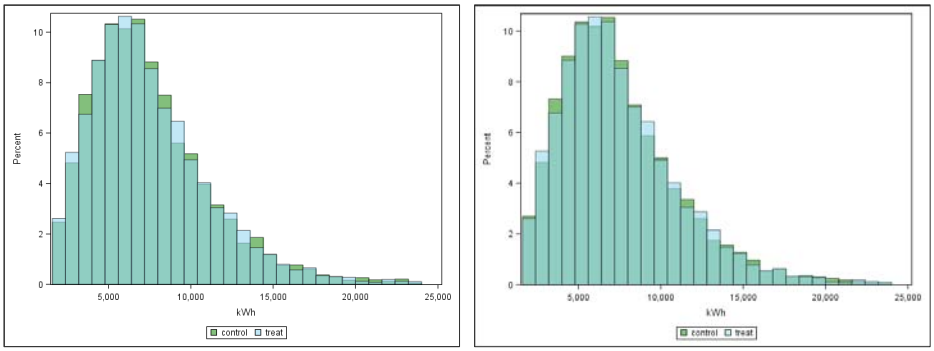


Figure 3: Consumption- and Model-Based AMI Data Matched kWh for Dual-Fuel Households



Second, we tested if the distributions of the matched samples are the same using the two-sample Kolmogorov-Smirnov (KS) test (see Test of Balance Section). We present the results from the test in Table 4. In general, the test results indicate samples whose distributions are not statistically different. With a probability value (p-value) of 0.05, we fail to reject the null that the sample data for SDG&E's opt-in and matched comparison groups we obtain using consumption-based matching and calendarized billing data come from the same population at the 95% confidence level. The p-value of 0.25 for SDG&E electricity data matched using model-based matching and calendarized billing data also indicates that we cannot reject that the null that matched comparison and opt-in samples have identical distributions.

The KS test results for both consumption- and model-based matching generated using AMI data also indicate well-balanced comparison and opt-in group matches. The high p-values of the tests for these matches suggest that the quality of matches improved substantially with AMI data. However, results from

[5] Plots for electric-only households are similar and we do not present them here to conserve space.

hypothesis tests of the kind KS test is applied to may be inapplicable in the current setting based on the Ho et al. (2007) and Imai et al. (2008) papers we discuss in the section on Test of Balance.

Table 4: Statistical test of balance for SDG&E matched electricity data

Matching technique	SDG&E electric	
	Test statistic	P-value
Billing data consumption-based matching	1.36	0.05
Billing data model-based matching	1.02	0.25
AMI data consumption-based matching	0.61	0.86
AMI data model-based matching	0.51	0.96

We present results from the standardized differences of propensity scores in Table 5. It is evident that both matching approaches using the two data types produce equally well-balanced matches. The standardized differences are all zero (to the second-decimal place). While we do observe that the standardized difference values drop the most from the unmatched to the matched data sets using AMI data, the final matched values for both data types and matching approaches indicate equally well-balanced matches.

Table 5: Standardized differences of propensity scores for SDG&E electricity data

Matching technique	Dual-fuel unmatched	Dual-fuel matched	Elec-only unmatched	Elec-only matched
Billing data consumption-based matching	0.047	0.000	0.079	0.004
Billing data model-based matching	0.057	0.000	0.093	0.003
AMI data consumption-based matching	0.079	0.000	0.090	0.001
AMI data model-based matching	0.072	0.000	0.096	0.002

Similarly, the ratios of the variances of the propensity scores are 1 or near 1, at the two-digit level, indicating that both methods and data types produce equally well-balanced matches (Table 6). While the matches using AMI data for electric-only households have values for this ratio that are marginally lower than those produced using billing data, these differences are very small and are not different enough to indicate superior matches using AMI data. We note that we start with data that are generally well-balanced (as we use matched comparison groups from the initial 5:1 matching exercise), and data type and matching approach do not seem to have material impact on matching outcomes.

Table 6: Ratios of propensity scores for SDG&E electricity data

Matching technique	Dual-fuel unmatched	Dual-fuel matched	Elec-only unmatched	Elec-only matched
Billing data consumption-based matching	0.97	1.00	1.18	1.09
Billing data model-based matching	0.99	1.00	0.88	1.05
AMI data consumption-based matching	0.96	1.00	2.62	1.02
AMI data model-based matching	0.94	1.00	0.97	1.02

Concluding Comments

Based on the results, using data from about 10,000 SDG&E electricity households that participated in a Universal Audit Tool opt-in program, we note that the two-step process that is customarily used by researchers to obtain best 1:1 matches using AMI data may be unnecessary. The 1:1 matches based on calendarized billing data are practically as good as matches based on the two-step

process that uses AMI data. Further, comparing balance using data from the same period where matching is done ('in-sample matching comparison') shows that model-based matching is of limited use. In future research, we will check if balance using data from a different period than the one used to do the matching ('out-of-sample matching comparison') may indicate that there are advantages to model-based matching.

We have also discovered the importance of the methods and metrics used to check balance in our study. The typically used hypothesis tests (and the one we use based on the KS test) may show improvements in balance where none exist. In our study, the KS test results indicate dramatic improvements in balance when we use AMI data in matching. They also indicate notable improvement in matching using model-based approaches. The propensity score diagnostics, however, do not indicate any such improvements, which highlights the importance of using an appropriate means of testing balance.

Additionally, visual tests of balance based on histograms and other plots of density while useful have some limitations. The distributions of the matched groups seem practically identical to the naked eye, yet propensity score diagnostics can pick small improvements in matches to the extent they exist.

References

Austin, P. C. and M.M. Mamdani. 2006. “A Comparison of Propensity Score Methods: A Case-study Estimating the Effectiveness of Post-ami Statin Use.” Statistics in Medicine 25: 2084–2106.

Ho, D.E., K. Imai, G. King, and E.A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Political Analysis 15:199–236.

Ho, D.E., K. Imai, G. King, and E.A. Stuart. 2011. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." Journal of Statistical Software 42 (8): 1-28.

Imai, K., G. King, and E.A. Stuart. 2008. “Misunderstandings Between Experimentalists and Observationalists About Causal Inference.” Journal of the Royal Statistical Society A 171 (2): 481–502.

Stuart, E.A. and D.B. Rubin. "Best Practices in Quasi-Experimental Designs: Matching methods for causal inference." In Best Practices in Quantitative Social Science, edited by J. Osborne, 155-176. Thousand Oaks, CA: Sage Publications, 2007.

Stuart, E.A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." Statistical Science 25 (1): 1-21.

Sawilowsky, S. 2009. "New effect size rules of thumb." Journal of Modern Applied Statistical Methods 8 (2): 467–474.