Losing Control: What Will Happen if Randomized Controlled Trials are Phased Out of Behavioral Program Evaluation?

Josh Schellenberg, Nexant, San Francisco, CA
Aimee Savage, Nexant, San Francisco, CA
Marshall Blundell, Nexant, New York, NY
Jonathan Cook, Nexant, Washington, DC
Brian Arthur Smith, Pacific Gas and Electric Co., San Francisco, CA[1]

ABSTRACT

Home energy reports and other behavioral conservation programs have gained significant traction in the utility industry as a supplement to traditional energy efficiency programs. The strongly preferred research design for evaluating these behavioral programs has been the randomized controlled trial (RCT), which typically relies on large control groups of non-participants. As these behavioral programs have grown from small pilots to full-scale programs that include a significant portion of the residential customer base, utilities and other stakeholders have started to question the future viability of the RCT, given that its large control groups of non-participants limit the potential for behavioral programs. Meanwhile, "Big Data" and innovative statistical methods from the fields of machine learning, Bayesian statistics and econometrics have also gained significant traction, especially in other industries such as the tech sector (Varian 2014). While these methodological alternatives show considerable promise for behavioral program evaluation, their statistical validity relative to the current "gold standard" – RCT – has yet to be tested rigorously. This paper leverages data from one of the largest home energy reports programs in order to compare the results of a large, multi-year RCT evaluation to the energy savings that evaluators would have estimated using alternative methods. Results from the analysis show that the alternative methods tested do not produce energy savings estimates that are similar to those of an RCT.

Introduction

As a complement to traditional rebate programs aimed at incentivizing energy-efficient purchases, behavioral approaches to generating energy savings have become a valuable part of the utility's energy efficiency (EE) portfolio. These "behavioral" programs aim to influence energy-using behavior and include home energy reports (HERs), building operator certification, gamification[2], information feedback, "pay-as-you-go" billing, enhanced marketing and strategies for increasing customer awareness (Friedrich, et al. 2010). Research from the behavioral sciences has suggested that such techniques are capable of achieving meaningful savings (see Delmas, Fischlein, & Asensio (2013) for a review), and these results combined with strong policy initiatives in many states have resulted in the rapid expansion of behavioral programs. ACEEE estimates that between 2008 and 2013, there were almost 300 behavioral programs being implemented across the US. This existing momentum and the prominent role of EE and conservation in EPA's Clean Power Plan proposal (EPA 2014) suggest that behavioral programs will continue to be important tools for meeting emissions reduction goals.

Like traditional EE programs, operationalizing behavioral programs requires measuring savings produced by a program in a robust manner. Conceptually, the savings attributable to a given program are

[1] The authors would like to thank Hunt Allcott from New York University, Alex Orfei from Opower, Ken Agnew from DNV GL, Peter Franzese from the California Public Utilities Commission (CPUC), and Dan Bush from the CPUC for their helpful feedback on which alternative methods to consider in this paper.

[2] Gamification is the use of the thinking and mechanics involved in structured play in non-game contexts to engage users.

equal to the difference between the (observed) consumption for customers who participate in the program and the (counterfactual) consumption for those same customers in the absence of the program. Because it is impossible to observe the same customers both participating and not participating in the program at the same time, consumption in the absence of the program must be estimated. This is the fundamental challenge of impact evaluations, which aim to measure the changes in energy usage that were caused by the program and not by other factors.

In the absence of information on the counterfactual, the next best solution is to construct a comparison group that is similar to the group of customers who participate in the program. The goal of choosing a comparison group is to eliminate selection bias, which occurs when there are differences in consumption between program participants and the comparison group *before* the start of the program. Ideally, the two groups would be very similar so that the consumption in the comparison group will provide good estimates of what would have happened to the program participants had they not participated. The gold standard for this approach for impact evaluations has been the randomized controlled trial (RCT), which uses random assignment to separate customers into a treatment group and a control group that does not participate. With proper implementation and a large enough sample[3], the use of random assignment ensures that the two groups will be equivalent prior to the onset of the program, which allows the counterfactual consumption of the treatment group to be estimated by observing the consumption of the control group.

Although RCTs are ideal in theory, there are several reasons why they may be infeasible or undesirable to conduct in practice. First, most require substantial resources (time, funding, statistical expertise, etc.) and so they may not be the best approach when quick answers are needed or when evaluation budgets are tight. Second, it may be that there are not enough customers who are both willing and eligible to participate. In this case, the small samples would reduce the precision of the impact estimate and potentially result in non-trivial differences between the treatment and control groups being undetected due insufficient statistical power. Third, implementing an RCT generally requires denying the treatment to a subset of the participants (the control group) for at least some period. If the treatment is thought to be beneficial, there can be political, legal or ethical concerns about withholding the treatment (Shaddish, Cook, & Campbell 2002; Khandker, Koolwal, & Samad 2013). Finally, as behavioral programs grow from small pilots to full-scale programs that include a significant portion of the residential customer base, utilities and other have started to question the future viability of the RCT due to concerns regarding the need for large control groups of non-participants limiting future rollouts.

In part due to these practical challenges, considerable research has been put into developing comparison group methods that do not utilize random assignment, but are still capable of providing valid impact estimates. These methods have been developed in a variety of different settings[4] and rely on techniques to eliminate selection bias and construct a comparison group of non-participants that is similar to the program participants. Quasi-experimental methods that have gained traction in the evaluation of EE and behavioral programs include propensity-score matching (PSM), regression discontinuity and explicit modeling of selection bias. Though there are instances of matching techniques being able to produce valid impact estimates in labor economics (Dehejia and Wahba 2002) and demand response applications (Nexant 2013), skepticism remains about whether these techniques can work in the context of EE and behavioral programs, especially those which produce relatively modest impacts ranging from 0.5% to 2.5% of whole-building usage (Allcott 2015).

[3] Behavioral programs can be operated on either an opt-in or opt-out (default) basis. RCTs are much easier to implement in a default program setting.

[4] Impact evaluation is a fundamental component of analysis in both the natural and social sciences. Comprehensive academic and industry literature on experimental and quasi-experimental methods can be found for psychology, healthcare, education, economics, public policy and other disciplines. A detailed review of this literature is beyond the scope of this paper.

In addition to the more traditional experimental and quasi-experimental evaluation methods, the recent proliferation of "big data" has provided yet another approach known as statistical learning.[5] As a new subfield of statistics, statistical learning is a set of techniques that traces its roots back to the 1970s and commonly abandons the concept of comparison groups altogether in favor of using advanced algorithms and large amounts of detailed pre-treatment data to predict the counterfactual outcome for the treatment group (James et al. 2013). These methods use a large amount of data to estimate more flexible relationships between variables than simple linear models, which may lead to more effective ways to model complex relationships involving energy usage. With the increased availability of large datasets from advanced metering infrastructure, using statistical learning techniques presents the opportunity to potentially no longer rely on large control groups of non-participants.

While quasi-experimental and statistical learning techniques show considerable promise for behavioral program evaluation, the statistical validity of each one relative to the RCT has yet to be tested rigorously in the context of measuring energy savings. The goal of this paper is to provide one such rigorous test by comparing impact estimates based on different methods for Opower's large-scale HER implementation at Pacific Gas and Electric Company (PG&E). Using the appropriate data for each method, savings estimates from the program are re-estimated using PSM, a Bayesian Structured Time Series (BSTS) model and a statistical learning model known as Regression Tree with Random Effects (RE-EM trees). The original RCT impact estimate is used as a benchmark for comparison and several robustness checks for each method (including the RCT) are explored.

Home Energy Report Programs

A growing number of companies offer programs to lower energy usage that feature energy usage feedback and behavioral suggestions. The most successful of these provides residential customers with periodic home energy reports (HERs) that compare a customer's monthly electric and/or gas usage to the average usage of similar homes as well as the average usage of a group of particularly efficient homes. Such "neighbor comparisons" can be based on a variety of customer characteristics, including location, home square-footage, presence of a pool or spa, and type and number of cooling and/or heating units.

Based on the neighbor comparisons, customers are given a rating along with additional information, such as a dollar amount of savings that the customer could realize on their energy bills by matching their efficient neighbors' usage or implementing behavioral and other changes. For customers receiving reports on their electric usage, the reports often include comparative graphs of their load shape.

The general hypothesis behind HERs is that neighbor comparisons provide a social motivation for customers to adjust their energy usage habits. Several studies of Opower HERs have a typical reduction of 0.5% to 2.5% on annual energy usage (Allcott 2015). This effect is thought to be primarily due to changes in behavior (turning off lights, adjusting the thermostat a few degrees, etc.), rather than significant investments in more efficient equipment. Interval data analysis provides evidence that savings begins within days of households receiving the first HER and identifies substantial savings resulting from changing thermostat settings (Todd et al 2014). Home inventories found that increased uptake of efficient lighting is a major energy savings driver (Perry and Woehleke 2013).

Data

In this comparative methods analysis, the approximately 75,000 treatment and 75,000 control customers who participated in one of the first HER RCTs at PG&E (the "Gamma" wave) are analyzed

[5] Statistical learning is sometimes called "advanced predictive modeling" and is closely related to (but not synonymous with) machine learning.

over the course of three post-treatment years (2012-2014). Importantly, HERs can only be sent to customers satisfying certain criteria, such as having a full year of bills, having a functioning SmartMeter for greater than one year and being on selected rate schedules. As a result, there are some differences in geographic distribution or average usage distribution between the eligible customers and the PG&E residential population.

For this comparative methods analysis, monthly electric billing data was provided by PG&E for all treatment customers and around 2 million non-participant households, including the RCT control group for the Gamma wave. The dataset includes one year of pre-treatment billing periods and data for all post-treatment months. Households are retained in each dataset until their accounts close. Because billing periods generally do not align with the calendar months, a month was assigned to each billing observation. The month of each bill is defined to be the month of the mid-date of the billing period.

Methods Tested

Measuring a 0.5% to 2.5% treatment effect accurately using whole-house electricity consumption data is a difficult statistical challenge. Individual household usage shows significant variation, and large sample sizes are needed to confidently distinguish any estimated savings from zero. The analysis presented here provides just one type of evaluation environment that is relevant for EE and behavioral programs. Though we can be confident that RCTs will perform well in other environments provided sample sizes are large enough, the relative ranking or effectiveness of the other methods may not necessarily be constant. In fact, it is likely that some set of empirical conditions exist for which each method is best suited. Examining these optimal conditions is beyond the scope of this paper, but this general point should be kept in mind whenever different methods are compared.

Randomized Controlled Trial (RCT)

PG&E's HER program lends itself well to an RCT because there is no recruiting process: it is an "opt-out" design whereby customers are simply assigned either to treatment or control groups according to inclusion criteria established at the onset of each experiment. Given the random assignment, the basic approach for estimating savings resulting is to simply compare the consumption of treatment and control customers. This was implemented for the PG&E evaluation using a panel regression model that included an indicator variable for month of the study, a treatment month indicator variable and a customer-level indicator variable (fixed effect) as explanatory variables. The regression equation is presented below.

$$kWh_{it} = customer_i \cdot b_i + \sum_{m=1}^{12} \sum_{y=2010}^{2014} I_{my} \cdot b_{my} + \sum_{j=1}^{n} \tau_j \cdot treatmonth_j + \varepsilon_{it}$$

This model is a generalization of a differences-in-differences methodology that improves the precision of the savings estimate by making use of pre-treatment data. In the above specification, the estimated savings for each month are given by the *treatmonth* parameter estimates. Average impacts for the post-treatment period can be obtained either by replacing the *treatmonth* variables with a single indicator variable equal to one for treatment customers in the post-treatment period, or by computing an average of the impacts in each month weighted by the number of treatment customers. Throughout the remainder of this paper, we treat the RCT results as the benchmark set of estimates to which the other methods are compared.

It is important to note that the size of the control group does not necessarily have to grow with the size of the treatment group in order to maintain the same level of statistical power. However, as programs grow to full-scale, both utilities and their regulators increasingly expect results to be segmented by several variables of interest such as climate zone, income, home owner/renter status, etc. Therefore, expanding the treatment group without a corresponding increase in the size of the control group may limit the flexibility to segment the results by all customer attributes that may be of interest.

Propensity Score Matching (PSM)

The goal of PSM is to create a control group in the absence of a randomized experiment that provides an accurate estimate of the treatment group's consumption had they not received the treatment. The first step is to estimate the probability of a customer participating in the program using a dataset containing participants and non-participants.[6] Each participating customer is then matched to a non-participant on the basis of this probability, or propensity score. Different approaches can be used to perform this matching, including nearest-neighbor matching, caliper and radius matching, stratification and interval matching and kernel matching (Khandker, Koolwal and Samad 2013). PSM relies on the variables included in the matching model being able to remove any selection bias that may exist. In other words, matching assumes that participation in the program can be explained by observed characteristics so that once those characteristics are accounted for, treatment is effectively random (as in an RCT).

PSM has a strong theoretical basis (Cameron and Triveldi 2005) and has been used to obtain impact estimates in a wide variety of applications. There also are select instances where estimates from matching have been compared to those from an RCT. Dehejia and Wahba (2002) conducted an analysis of the National Supported Work (NSW) Demonstration[7] using PSM and found that the matching estimates compared very well with the experimental benchmark. Nexant has performed similar comparisons for demand response impacts (Nexant 2013) and also found that matching produced similar results to a pure experimental approach.

To analyze the HER program using matching, we developed a matched control group from a large dataset of non-participants (approximately 2 million customers were eligible to be matched to a treatment customer). Each of the 75,000 treatment customers was matched to a customer in the large dataset of non-participants using PSM. The variables included in the propensity score model were simply the pretreatment monthly usage amounts for November 2010 through October 2011. An additional constraint was that each treatment customer had to be matched to a non-participant from within the same weather station area. This constraint ensured that matches were drawn from within the same geographic area. Figure 1 provides a comparison of pre-treatment usage for treatment customers and the matched control group. From November 2010 through October 2011 (the months used in the match), the difference between treatment usage and matched control group usage is very small (the discrepancy does not exceed 0.3% in any month). As a result, the two lines representing pre-treatment usage for the two groups match so closely that the lines are indistinguishable in the graph.

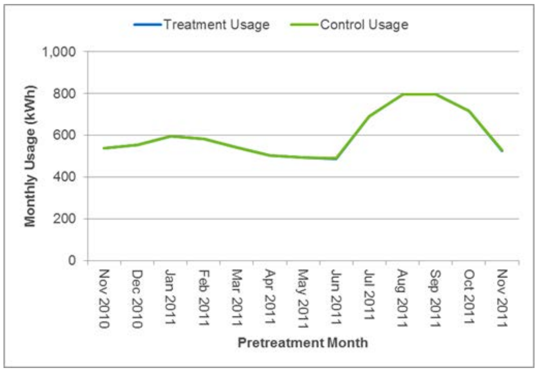

Figure 1. Monthly Usage for Treatment and Matched Control Group during Pre-treatment[8]

[6] Logit and probit models are typically used for this application.

[7] The NSW Demonstration was a labor market experiment designed to test whether work training programs would improve the ability of hard-to-employ people to get and hold normal, unsubsidized jobs. For additional background, refer to: http://www.mdrc.org/publication/summary-and-findings-national-supported-work-demonstration.

[8] Treatment usage and control usage match so closely that the lines are indistinguishable in the graph.

Bayesian Structured Time Series (BSTS)

One example of a statistical learning model that is capable of modeling counterfactual electricity consumption is a BSTS model. BSTS models aim to create a “synthetic control” (Abadie, Diamond and Hainmueller 2010) by making use of three distinct sources of information: trends in electricity consumption before the start of a program, trends in other variables that are related to electricity consumption before the start of the program and prior knowledge about model parameters that may be available from previous studies. By fitting the model to the pre-treatment period, the resulting parameter estimates can be used to predict counterfactual consumption in the post-treatment period given the pre-treatment time series and post-treatment values for the covariates included in the model. Once estimated, the impact estimate for the program is calculated simply as the difference between the observed consumption and the predicted consumption that would have occurred had the program not existed.

BSTS models can be an appealing option in cases where the requirements for an RCT are not met. One attractive feature of the approach is that it provides considerable flexibility in modeling trend, seasonality and regression effects that allow for complex, non-linear relationships between variables (Scott and Varian 2014). It also offers potential improvements over classic difference-in-difference methods by providing a full time-series estimate for the effect of a program and allowing for model averaging to select the most appropriate synthetic control. The primary disadvantages of BSTS models are their relative complexity (especially compared to randomized experiments) and the modeling expertise required to successfully implement them. Statistical learning models, in general, come with a risk of “overfitting”, whereby too much importance is placed on random patterns in the data, which can lead to poor out-of-sample predictions (Kuhn and Johnson, 2013).

Brodersen, et al. (2014) developed a BSTS model and used it to analyze a Google advertising campaign in the United States. The impacts of the advertising campaign had previously been analyzed using a randomized experiment, which allowed the authors to compare the results of the BSTS model to those obtained from the experiment. Results from the analysis showed that explicitly modeling the counterfactual consumption of the treatment group produced nearly the same estimates as the experiment, but with slightly less precision. Another output from the Brodersen study was a package for the statistical program `R` called `CausalImpact`, which provides an implementation of the model.

To apply the BSTS model, the `CausalImpact` package was obtained online (http://google.github.io/CausalImpact/) and applied to an analysis dataset consisting of average kWh for the treatment group, heating degree days (HDD), cooling degree days (CDD), relative humidity, and their higher powers and interactions. We also include a monthly seasonal effect. A critical assumption to obtain a valid causal inference is that the covariates that serve as a synthetic control were unaffected by the intervention, which certainly holds in this case. While BSTS can be combined with data from a randomly selected control group, we do not include covariates associated with untreated customers, as we wish to test the viability of the method to estimate savings of large scale programs for which control participants may not be available.

Regression Tree with Random Effects (RE-EM Trees)

RE-EM trees are an application of a tree-based estimation method to longitudinal data. They combine the flexibility of tree-based methods with the structure of random effects models. Mixed effects models are a common way to control for unobserved heterogeneity, assuming that the individual specific effects are uncorrelated with the independent variables. Tree-based models are rule-based models that partition data based on one or more nested *if-then* statements applied to the independent variables. The predictor space is cut into regions and the outcome is then predicted by a single number within each region (Kuhn and Johnson, 2013).

Tree-based models are popular for their ease of interpretability and implementation. The models can handle many types of predictors without any data pre-processing, and generally make no assumptions about the data structure (the RE-EM tree makes the random effects model assumptions). The models can also accommodate missing data and implicitly conduct feature selection. However, tree based models are prone to model instability, and may exhibit poor predictive performance if the relationship between predictors and response cannot be adequately defined by rectangular subspaces of the predictors (Kuhn and Johnson, 2013).

Sela and Simonoff (2012) developed the RE-EM tree estimation method and applied it to pricing online transactions. They showed that the RE-EM tree method is relatively less sensitive to parametric assumptions and provides improved predictive performance compared to linear models with random effects and regression trees without random effects. They also apply it to a dataset examining traffic fatalities and show that the RE-EM tree strongly outperforms a tree without random effects and performs comparably to a linear model with random effects.

To estimate energy savings from the HER program using a RE-EM tree model, we modeled monthly energy consumption in kWh of treatment customers in the pretreatment period of November 2010 through October 2011. Predictors included average monthly relative humidity, HDD, CDD, a categorical variable for month of the year, as well as whether the customer was on the California Alternative Rates for Energy (CARE) program for low-income customers and whether they became a net energy metered (NEM) customer at some point during the post-treatment period. In order to include a categorical variable encoding month as a predictor, at least a year of data was needed to fit the model. This left only one month of energy usage available for cross-validation (October 2011), so different parameters that control the fit of the model were not tuned, and the default values were used. Once a fitted model is obtained, it is used to predict energy consumption in the post-treatment period for each customer. Those predictions serve as the counterfactual from which actual energy consumption is subtracted to estimate savings.

Results

Randomized Controlled Trial (RCT)

The RCT analysis dataset contains one year of pre-treatment data (2011) and three years of post-treatment data (2012-2014) for around 150,000 customers, randomly divided evenly between treatment and control. Average annual percent impacts for Gamma wave customers who receive the standard report frequency of HERs[9] in each of the three post-treatment years are presented in Table 1. Estimates indicate a gradual increase in savings from year to year, starting at around 1.2% in 2012 and increasing to nearly 1.7% in 2014. The general magnitude of these percent savings is consistent with the typical impact of Opower HERs, which ranges between 0.5% and 2.5%.

Table 1: Average Annual Percent Energy Savings for 2012-2014 using the RCT

Year	Avg. % Impact
2012	1.16%
2013	1.58%
2014	1.69%

In addition to the average annual impacts, month-specific savings were estimated in order to observe the trend in treatment effects throughout the year. Table 2 presents the average kWh and percentage impact by month in 2014. Savings vary throughout the year, with absolute kWh savings peaking in the summer and winter. On a percentage basis, the impact is nearly 1.7% on average, but ranges from a low of just over 1%

[9] The Gamma wave of HERs is separated into dual-fuel “standard report frequency,” dual-fuel “reduced report frequency,” and electric-only customers. This stratification allows for the comparison of the frequency of HERs on energy usage as well as the effect on customers with different fuel-types. This paper focuses on the standard report frequency group.

in the summer months to a high of a little more than 2% during the winter months. A similar trend is found in the monthly analysis for 2012 and 2013, but those years are omitted here for brevity.

Table 2: Average 2014 Monthly Energy Savings using the RCT

Month	Avg. Impact (kWh)	% Impact	Std. Error (kWh)
Jan 2014	11.8	2.2%	1.4
Feb 2014	10.6	2.1%	1.4
Mar 2014	9.9	2.2%	1.3
Apr 2014	9.4	2.0%	1.2
May 2014	10.1	1.7%	1.4
Jun 2014	11.3	1.5%	2.0
Jul 2014	10.4	1.2%	2.5
Aug 2014	10.3	1.3%	2.3
Sep 2014	9.9	1.5%	1.7
Oct 2014	9.5	1.9%	1.3
Nov 2014	9.6	2.0%	1.4
Dec 2014	8.6	1.6%	1.6

As in the absolute and relative impacts, the precision of the estimates also changes during the course of the year. The most precise estimates are obtained during the spring and fall months, when HVAC usage is low and overall usage is less variable. The least precise estimates occur for the summer months when differences in climate, thermal preferences and cooling equipment (both A/C and thermostats) combine to result in more variation in usage at the individual customer level. While the standard errors could likely have been reduced by including additional covariates to control for variation in weather in the model, we elected not to do so. One of the advantages of an RCT estimated using difference-in-differences is that it does not require performing model selection and so does not suffer from model selection bias or specification error.

Propensity Score Matching (PSM)

Table 3 summarizes the annual percent energy savings estimates for the PSM model as compared to the RCT results. In general, the two sets of percent savings estimates are somewhat close, particularly in the first year. Nonetheless, it is important to recognize that for small treatment effects like those that arise from HERs, a seemingly small discrepancy between two percent impact estimates may ultimately correspond with a large difference in the estimated energy savings.

Table 3: Average Annual Percent Energy Savings for 2012-2014 using PSM

Year	Avg. % Impact Estimate	
	RCT	PSM
2012	1.16%	2.08%
2013	1.58%	3.07%
2014	1.69%	3.21%

In order to directly compare the kWh savings estimates from the two models, Figure 2 shows the monthly savings estimates using PSM as compared to the RCT results from 2012 through 2014. Although the percent savings estimates are relatively close and the pre-treatment usage for the matched control group suggests that it represents the treatment group as well as the RCT control group, the resulting kWh savings estimates are quite different from those of the RCT. In the first post-treatment month, the savings estimates are already nearly 50% higher for the PSM model as compared to the RCT estimates. This upward bias in the savings estimates persists throughout all post-treatment months. For the post-treatment period overall, the upward bias in the PSM results is 89%.

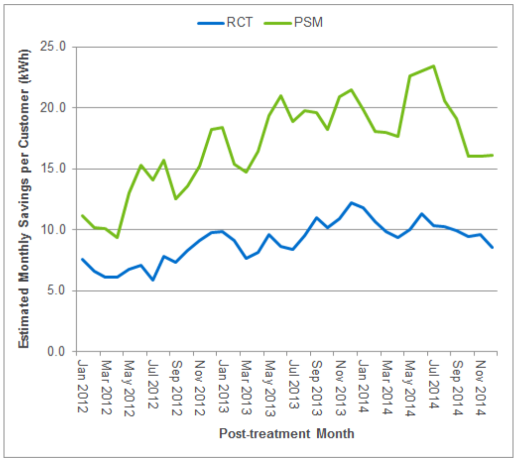

Figure 2: Estimated Monthly kWh Savings per Customer using PSM

Bayesian Structured Time Series (BSTS)

Table 4 summarizes the annual percent energy savings estimates for the BSTS model as compared to RCT. The BSTS savings estimates are highly variable from year to year and are quite different from the RCT results. In 2014, the BSTS percent savings estimate is nearly four times higher than the RCT estimate.

Table 4: Average Annual Percent Energy Savings for 2012-2014 using BSTS

Year	Avg. % Impact Estimate	
	RCT	BSTS
2012	1.16%	-0.40%
2013	1.58%	2.21%
2014	1.69%	6.43%

Figure 3 shows the monthly savings estimates using BSTS as compared to the RCT results from 2012 through 2014. The resulting savings estimates are different from the RCT results in both their general magnitude and monthly variation.

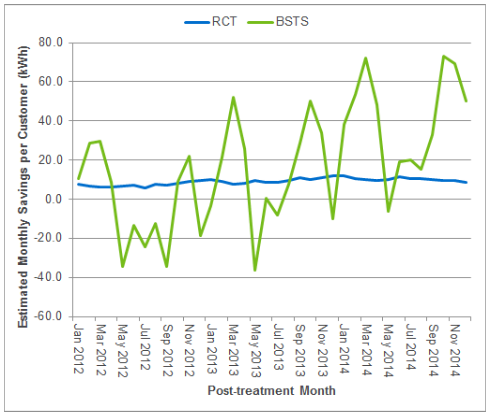

Figure 3: Estimated Monthly kWh Savings per Customer using BSTS

The BSTS per-customer savings estimates are noisy throughout all post-treatment months, varying from as low as *negative* 36.4 kWh to as high as 73.1 kWh. This range in the BSTS per-customer savings estimates is substantially wider than the range of 5.9 kWh to 12.2 kWh in the monthly RCT estimates.

Regression Tree with Random Effects (RE-EM Trees)

Table 5 summarizes the annual percent energy savings estimates for the RE-EM tree model as compared to the RCT results. The RE-EM tree percent savings estimates are somewhat similar in the first year, but then diverge substantially from the RCT results over time. As in the BSTS results, the 2014 RE-EM tree percent savings estimate is nearly four times higher than the 2014 RCT estimate.

Table 5: Average Annual Percent Energy Savings for 2012-2014 using RE-EM Tree

Year	Avg. % Impact Estimate	
	RCT	RE-EM Tree
2012	1.16%	2.03%
2013	1.58%	3.07%
2014	1.69%	6.08%

Figure 4 shows the monthly savings estimates using the RE-EM tree model as compared to the RCT results. Similar to the BSTS findings, the RE-EM monthly savings estimates are completely different from the RCT results in both their general magnitude and variation from month to month.

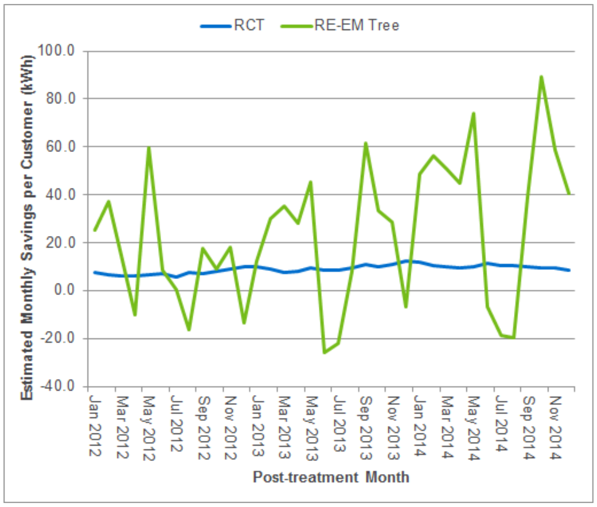

Figure 4: Estimated Monthly kWh Savings per Customer using RE-EM Tree

In the first post-treatment month, the savings estimates are already over three times as high for the RE-EM tree model as compared to the RCT estimates. The RE-EM tree per-customer savings estimates are noisy throughout all post-treatment months, varying from as low as *negative* 25.8 kWh to as high as 89.3 kWh. This range in the RE-EM tree per-customer savings estimates is substantially wider than the range of 5.9 kWh to 12.2 kWh in the monthly RCT estimates. The RE-EM tree estimates also tend to increase and become extremely variable in later months.

Conclusion

This paper leverages data from one of the largest HERs programs to compare the results of an RCT to estimates using alternative methods. While producing results that are similar to those of a RCT is difficult in most circumstances, the modeling task in this case is particularly challenging because the treatment effect comprises a relatively small percentage of whole-house usage. On an annual basis, the RCT produces percent savings estimates that range from 1.16% in 2012 to 1.69% in 2014. The RCT monthly energy savings estimates are fairly stable, varying from 5.9 kWh to 12.2 kWh per customer over 36 monthly estimates. As shown in Table 6, the estimates from the three alternative methods tested are different. In particular, the results for the BSTS and RE-EM tree models are highly variable from month to month, with relatively large *negative* savings estimates in some months and extremely high estimates in others. Among the three alternative methods tested on this data, PSM is the most viable alternative to the RCT. However, it is important to note that the PSM analysis in this paper relied on a large pool of around 2 million non-participants from which to draw the matched control group. As with the RCT, the PSM approach does not resolve the issue of requiring a large comparison group of customers who do not receive the treatment.

Table 6: Summary of Comparative Methods Analysis

Estimation Methodology	Percent Savings Estimates			Monthly Savings Estimates (kWh)	
	2012	2013	2014	Low	High
Randomized Controlled Trial (RCT)	1.16%	1.58%	1.69%	5.9	12.2
Propensity Score Matching (PSM)	2.08%	3.07%	3.21%	9.4	23.4
Bayesian Structured Time Series (BSTS)	-0.40%	2.21%	6.43%	-36.4	73.1
Regression Tree with Random Effects (RE-EM Tree)	2.03%	3.07%	6.08%	-25.8	89.3

Both the BSTS and RE-EM tree results show a large increase in estimated percent savings in 2014. This false effect is likely attributable to changes in usage and weather conditions that occurred from 2013 to 2014; average consumption per customer decreased by 1.8% from 2013 to 2014, while the average temperature increased. This led to a large upward bias in the 2014 BSTS and RE-EM estimates, given that both models primarily rely on temperature to estimate usage. Another limitation of the BSTS and RE-EM tree models tested in this paper is that they forecast usage over multiple years having been trained on only 13 months of pre-treatment data, which may explain why the estimates from these models are so variable from month to month. Further research based on several years of hourly interval data is required to conclusively determine whether these models are (or are not) a viable alternative to the RCT. Nonetheless, this analysis shows that a model that primarily relies on temperature patterns as a predictor of usage, whether it be monthly or hourly usage, may go awry after several years of treatment if there are fundamental changes in customer usage over time that are not due to temperature. Among the alternative methods tested in this paper, the key advantage of the PSM approach is that it does not rely on modeling a relationship between temperature and usage, which most likely explains why the PSM results track most closely to the RCT results over multiple years.

References

Abadie, Alberto, Diamond, A. and Hainmueller, J. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program." *Journal of the American Statistical Association* 105.490.

Allcott, Hunt. 2015 (forthcoming). “Site Selection Bias in Program Evaluation.” *The Quarterly Journal of Economics* 130:3.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. 2014. “Inferring causal impact using Bayesian structural time-series models.” *The Annals of Applied Statistics*.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: methods and applications.* Cambridge university press.

Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity score-matching methods for nonexperimental causal studies." *Review of Economics and Statistics* 84.1: 151-161.

Delmas, Magali A., Miriam Fischlein, and Omar I. Asensio. 2013. "Information strategies and energy conservation behavior: A meta-analysis of experimental studies from 1975 to 2012." *Energy Policy* 61: 729-739.

Environmental Protection Agency (EPA). 2014, June 18. *Carbon Pollution Emission Guidelines for Existing Stationary Sources: Electric Utility Generating Units*. Retrieved from the Federal Register website: http://www.gpo.gov/fdsys/pkg/FR-2014-06-18/pdf/2014-13726.pdf. (79 FR 34829).

Friedrich, K., Amann, J., Vaidyanathan, S., & Elliott, R. N. 2010. “Visible and concrete savings: Case studies of effective behavioral approaches to improving customer energy efficiency.” American Council for an Energy-Efficient Economy.

James, G., Witten, D., Hastie, T., & Tibshirani, R. 2013. *An Introduction to Statistical Learning*. New York: Springer.

Khandker, Shahidur R., Gayatri B. Koolwal, and Hussain A. Samad. 2013. *Handbook on Impact Evaluation: Quantitative Methods and Practices.* World Bank Publications.

Kuhn, M., & Johnson, K. 2013. *Applied predictive modeling*. New York: Springer.

Nexant. 2013. *SmartPricing Options Interim Evaluation.*

Perry, Michael and Woehleke, Sarah 2013. “Evaluation of Pacific Gas and Electric Company's Home Energy Report Initiative for the 2010–2012 Program.” CALMAC ID PGE0329.01.

Scott, Steven L., and Varian, Hal R. 2014. "Predicting the present with Bayesian structural time series." *International Journal of Mathematical Modelling and Numerical Optimisation* 5.1: 4-23.

Sela, Rebecca J., and Jeffrey S. Simonoff, 2012. “RE-EM trees: a data mining approach for longitudinal and clustered data.” *Machine Learning* 86: 167-207.

Todd, Annika, Michael Perry, Brian Smith, Michael J. Sullivan, Peter Cappers, and Charles A. Goldman, 2014. *Insights from Smart Meters: The Potential for Peak-Hour Savings from Behavior-Based Programs.* http://emp.lbl.gov/publications/insights-smart-meters-potential-peak-hour-savings-behavior-based-programs. Berkeley, Calif.: Lawrence Berkeley National Laboratory.

Varian, H. R. 2014. “Big data: New tricks for econometrics.” *Journal of Economic Perspectives,* 3-27.