Project: Standardized, Sustainable and Transparent EM&V - Integrating New Approaches

Subtask 1B.1 / 1B.2

Memo on Residential EM&V Methods and Technical Approaches Used by Residential Advanced M&V Tools

Samir Touzani, Jessica Granderson, Eliot Crowe

Building Technology and Urban Systems Division

Lawrence Berkeley National Laboratory

Prepared for:

Michele Melley

Connecticut Department of Energy and Environmental Protection

July 15, 2019

This memo is submitted with respect to Subtask 1A.3, Deliverable 1B.2.1, under U.S. Department of Energy Statement of Project Objectives (SOPO) DE-EE0007779/0000.

Tasks/Deliverables Description
Subtask 1B.1: Research Technical Methods Used in Automated Residential Tools (Months 19 – 27)
Subtask Summary: CT utilities, with vendors and support from LBNL, will investigate the technical approach used in today's automated residential M&V tools, and determine which elements of the savings evaluation methods in these tools are program- or application-dependent and which are constant, considering, for example, selection of the comparison group, correction for bias, determination of sample sizes, etc. Further, the pilot will address how automated consumption data analysis can be integrated or supplemented by evaluation to provide final evaluated results (i.e., for claiming savings to a regulatory commission) – i.e., 'EM&V 2.0'. The pilot design will also be informed by experience from the Commercial M&V 2.0 pilot, recognizing there will be both similarities and differences in methodology.

Subtask 1B.2: Assess Residential Tools with Respect to Current Practice in Residential EM&V (Months 22 – 27)
Subtask Summary: Drawing upon recent LBNL evaluation of data from 10 RCT utility time-based pricing pilots through the Smart Grid Investment Grant Consumer Behavior Studies, and related work in comparing "Gold Standard" RCT methods to quasi-experimental evaluation approaches, LBNL will qualitatively compare and contrast the approaches used in the residential automated M&V tools to other industry-standard methods and best practices. Based on 1B.1 and 1B.2 subtasks, and any other information gathered during this period, prepare residential pilot design and implementation plan.

Milestone 1B.2.1: Memo on research and findings from subtasks 1B.1 and 1B.2.

1. Introduction

Advanced M&V (sometimes called M&V 2.0 or automated M&V), is characterized by (1) increased data availability, primarily in terms of finer time scales or higher volume and (2) enabling the processing of large volumes of data at high speed via automated analytics, to give near real-time savings estimates. These approaches are intended to be conducted more quickly, more accurately, and potentially at lower cost than non-automated methods. For residential applications, advanced M&V typically employs monthly data across a high number of homes. Advanced M&V for residential programs offers utilities the possibility of getting an early look at actual savings, in advance of formal program evaluations that may not be completed for several years after the program completes. With this data, utilities have a chance to address potential issues soon after they arise.

Several previous studies (Granderson et al. 2016; Granderson et al. 2017) have analyzed the effectiveness and the accuracy of advanced M&V tools at the premise level in the commercial sector, but little has been published on the potential of these methods in evaluating energy efficiency (EE) programs in the residential sector[1]. Further, there is a need to understand how the techniques employed by advanced M&V tools for residential applications compare to established program evaluation methods.

This memo provides a brief summary of the main concepts and approaches that are used for residential EM&V. It will also summarize the technical approaches used in today's automated residential M&V tools, so that these technical approaches may be compared to current evaluation practices.

2. Summary of current concepts in residential EM&V

The following subsection is primarily drawn from the technical concepts presented in Agnew and Goldberg 2017, Violette and Rathbun 2017, and SEE Action 2012.

2.1. Energy savings

Energy savings from an EE program are defined as the difference between the quantity of energy consumed by the participants in the program and the energy that these participants would have used during the same period of time, if they had not participated in the program. However, it is impossible to measure the energy that participants *didn't* consume due to their participation in the program. Thus, one cannot "measure" energy savings, but it can be estimated.

A common savings estimation approach is to estimate the difference between the participants' energy use and the energy use of a comparison group that is considered to be drawn from a population of customers that are similar to the participants. A comparison group provides a proxy counterfactual that is used to estimate the impact of the EE program. It also aims to control to some extent for exogenous factors (exogenous to the EE program) that impact the energy consumption and that affect all the customers but are not related to the EE program measures (e.g., economy evolution, energy price changes, etc.). Depending on types of customers, the effects of these exogenous factors are different. Therefore, it is important that the customers of the comparison group are as similar as possible to the customers that participate in the EE program.

The difference in the energy use between the program participants and the comparison group can be due to three main components:

1. The impact of the EE program;
2. The selection bias that is due to the differences (in term of customers' characteristics) between the comparison group and the participant group;
3. The inherent randomness in how homes consume energy (also called random errors).

To obtain an accurate estimate of energy savings that are due to the impact of the EE program it is important to use approaches that minimize or eliminate the second and third component. Usually, the level of selection bias depends on how similar the comparison group is to the group of

[1] One example presentation of a residential advanced M&V tool evaluation was shared at a NEEP event, and can be downloaded at: https://neep.org/sites/default/files/EM%26V%20Fall%20Meeting_Calling%20All%20Pilots.pdf

participants. Thus, to obtain an unbiased energy savings estimate it is required to design a comparison group that is statistically identical to the participants group. The inherent randomness can be caused by unobservable and unpredictable factors that can impact the energy use (e.g., changes in customer behaviors, problem with metering, etc.). The impact of this type of uncertainty can be reduced by increasing the number of customers in both the comparison group and the participant group.

2.2. Gross vs. Net energy savings

Depending on the program specification the energy savings can be reported as gross savings and/or net savings values. The distinction between these values relates to the need to identify if the energy savings are due to the program (also known as “attribution”) or to other factors. These two metrics are defined as follow (Violette and Rathbun 2017):

- *Gross energy savings*: “The difference in energy consumption with the energy-efficiency measures promoted by the program in place versus what consumption would have been without those measures in place”. In other words, the energy use change within the participant group that is directly due to the EE measures, regardless of the motivation of their participation in the program.
- *Net energy savings*: “The difference in energy consumption with the program in place versus what consumption would have been without the program in place”. This energy savings metric typically requires considering two additional factors beyond gross savings: free-ridership and spillover:
  - *Free-ridership*: “the portion of energy savings that participants would have achieved in the absence of the program through their own initiatives and expenditures” (SEE Action 2012). In other words, participants that would have implemented the measures that are promoted by the program in the absence of the program (for more information refer to Violette and Rathbun 2017).
  - *Spillover*: “program-induced adoption of measures by nonparticipants and participants who did not claim financial or technical assistance for additional installations of measures supported by the program” (SEE Action 2012). For example, a participant can implement more EE measures as a result of the program, a non-participant may install EE measures as a result of exposure to the program, and in neither case are the resultant savings captured in program reporting (for more information refer to Violette and Rathbun 2017).

2.3. Comparison Group Methods

There are several methods for designing a comparison group that is statistically as similar as possible to the participant group and is suited to residential programs that have a relatively high number of participants with similar characteristics:

- *Future participants:* using customers that will participate in the EE program in the future as a comparison group to the customers that are participating in the current year. The two groups should be statistically similar if the participants mix is homogenous. In addition, the future participants have the same tendency to participate in the EE program as the customers in the participant group, which is reducing the self-selection bias.

- *Past participants:* using customers that previously participated in programs as a comparison group (similar characteristics as future participants approach).
- Randomized Controlled Trial: described in Section 2.3.1 below.
- Quasi-experimental methods: described in Section 2.3.2 below.

If using the first two types of comparison group listed above, the EE program needs to be running for multiple years with sufficient data collection over the analysis period. It should also be noted that the approach using "future participants" as a comparison group would, by definition, enforce a time lag (i.e., You need to wait until enough of those participants have signed up for the program), thereby defeating one of the key selling points of advanced M&V (i.e., the ability to assess savings in near real-time). In the following two sections we will discuss the randomized controlled trial and the quasi-experimental methods.

2.3.1. Randomized Controlled Trial (RCT)

The RCT approach is the 'gold standard' for designing comparison groups because it typically provides less biased and more precise results (i.e., smaller inherent randomness) than other approaches. However, RCT is not practical to implement for most EE programs. The RCT approach is defined as follows: before starting the program engagement, a study population is defined (e.g., eligible homes, geographic location, etc.). Then the customers are randomly assigned either to the comparison group or the treatment group (i.e., program participants). The energy consumption data must be gathered for all the customers within the treatment and the comparison group. This data can be either monthly bill data or metered data.

The random assignment removes the effects of observable differences such as the house characteristics (e.g., number of stories, floor area, etc.) and non-observable differences such as customer behavior (e.g., change in occupancy or energy consumption etc.). Therefore, if the RCT is correctly designed, it can produce a comparison group that is statistically similar to the treatment group, which eliminates selection bias.

If the net savings is defined as the gross savings plus the participants' spillover minus the free-ridership, then the RCT approach provide a reliable assessment of the net impacts of the program (i.e., net savings). The participant spillover is by definition captured and free-ridership is addressed by the fact that the comparison group and the treatment group will in theory contain a similar number of free-riders. Therefore, the energy savings of the free-ridership in the comparison group will cancel out the energy savings of the free-ridership in the treatment group. Note that while spillover within the participants' group can be addressed by RCT design, it will not capture spillover effects for non-participants. To appropriately address the non-participant spillover effect the evaluator needs to conduct an independent study of the comparison group customers (Violette and Rathbun 2017).

A well-designed RCT produces accurate net savings of the evaluated programs, because it minimizes the selection bias, and controls for free ridership and participant spillover. However, RCT is not practical for the majority of EE program evaluations for several reasons, for example the RCT approach involves planning in advance of the program implementation which is often impractical for program evaluations. As pointed out in (Violette and Rathbun 2017) it is also possible that random assignment can allocate customers to a participant group even if they end up not needing or not wanting the efficiency measures; There are also some equity issues associated with assigning some customers to the comparison group and thus not allowing them to benefit from an EE program funded by ratepayer dollars.

2.3.2. Quasi-experimental methods

In situations where an RCT approach is not feasible the evaluation can be performed using non-randomized methods to define comparison groups. These methods are called quasi-experimental methods (QEM) and are widely used in different fields such as medicine, economics and political science (Stuart 2010). Typically, in a QEM approach the customers self-select as participants, and the evaluation analysis needs to design the comparison group. Therefore, since the participant group is not randomly selected, the QEM approach is more vulnerable to selection bias that can lead to a biased estimation of the energy savings. In addition, if the comparison group is selected from the same pool of customers who were eligible to participate but chose not to, the participant and comparison groups may have some differences regarding their approach to their energy consumption that can lead to a significant bias in the energy savings estimate. There are several techniques to develop a comparison group for the QEM approach. The most common are the pre-post energy use method and the matching methods, which are described below.

Pre-post energy use methods

The main idea behind this method is to use the participants themselves as the comparison group. The energy consumption of the customers that are in the participants group is compared to their energy consumption prior to the efficiency measures installation (promoted by the program). The most challenging aspect of this approach is that there are several factors that can impact the change in the energy use before and after measure installation. Not accounting for these impacts will result in an unreliable estimation of the energy savings due to the program. Therefore, a simple comparison between the pre and post energy use is not a viable solution. Statistical regression analysis is used to improve the accuracy by minimizing the bias; however, it is not always possible to include all the influential factors as independent variable of the model. This is due to the fact that some of the factors can be accurately measured and widely available (e.g., outdoor air temperature, Zip code), some of them are more challenging to obtain (e.g., occupancy) and some are very difficult to observe and to account for (e.g., economy, change in behavior).

Matched comparison groups

In this approach the goal is to construct a (non-random) comparison group with customers selected from a pool of non-participants in the EE program, and that are as similar as possible to the participant group. The biggest challenge in using this approach is the availability of observable characteristics of the customers (e.g., number of occupants, income) and their houses (e.g., square footage, location) that can be used as matching factors. There are also non-observable factors that can affect the energy use and which are very hard or impossible to match (e.g., human behavior).

One approach for matching is based on grouping (i.e., stratifying) the customers based on some predefined characteristics (e.g., house square footage, zip code, average energy use), and then from the pool of non-participants randomly select the comparison group for each stratification level. This method requires the evaluator to predefine the matching characteristics and how the stratification levels are defined. Usually the number of non-participants within a stratification level is proportional to the number of participants within the same level.

Another approach matches each participant to a specific non-participant customer. This type of matching methods usually has four major steps (Stuart 2010; Violette and Rathbun 2017):

1. Define an appropriate similarity, which is the metric that measures if a non-participant customer is a good match to a participant.
2. Use the similarity metric to create the comparison group.
3. Asses the quality of the created comparison group, and if needed reiterate step 1 and 2
4. Estimate the impact of the EE program (i.e., estimate savings)

Defining the similarity metric involves consideration of two points: the first point is which variables (i.e., matching factors) will be used to define the similarity (variables that impact the energy use), and the second is defining the distance metric. Depending on their availability, the following features can be used as variables in defining similarity:

- Energy consumption in the pre-installation period (most commonly available);
- Vintage of the house;
- Square footage;
- Zip code (usually available);
- Number of bedrooms;
- Number of occupants;
- Income of the household.

There are several metrics that can be used to measure the distance (i.e., the similarity) between customers, for example the Euclidian distance, the Mahalanobis distance, and the propensity score. The latter of these usually involves using a logistic regression (note that a machine learning approach can also be used and often shows good results, see [Stuart 2010] for more details). The propensity estimation regression model will use the selected variables that will define the customers (participants and non-participants) and a dependent variable defined as equal to 1 if the customer is a participant and 0 if otherwise (i.e., propensity score). This model will allow for the identification of non-participant customers that have similar characteristics. For more information about the matching methods in general and the propensity score in particular we refer the reader to (Violette and Rathbun 2017; and Stuart 2010).

To summarize, “quasi-experimental designs try to replicate designs that employ randomization using observational (nonrandomized) data” (Violette and Rathbun 2017). This approach is easier to conduct than the RCT approach and can be initiated after program implementation, which is not the case for the RCT. The selection bias can be limited by a well-designed matching method, which can also partially control the free-ridership and the participant spillover. The QEM is also a common approach used by several scientific communities where the RCT is not possible, and thus can be referenced in extensive literature. However, if there are unobservable (or hard to measure) variables that significantly impact the energy consumption, which is usually the case, it is difficult to design a good comparison group using the matching methods, thus resulting in biased savings estimates. Moreover, there is the issue of self-selection in the program, and as pointed out in (Agnew and Goldberg 2017): "A key characteristic on which we'd like the comparison group to match the participants is whether the customer would adopt the energy efficiency activity in the absence of the program", which is affecting the ability of the QEM to estimate the net savings. Residential evaluation using comparison groups will [a] almost always involve compromising between practicality/cost and accuracy; [b] include some level of error and/or bias that cannot be quantified; and [c] be subject to the expertise and judgement of the evaluator in suiting the evaluation method to the program and available data.

3. M&V techniques in advanced M&V tools

Most residential advanced M&V software tools that are available (Granderson and Fernandes 2017) employ IPMVP Option C approach. The metered data can be either monthly energy bill data or more granular (e.g., daily, hourly, etc.). A statistical regression model (i.e., baseline model) is created for each analyzed building in order to estimate the saved energy. These models usually involve outdoor air temperature as an independent variable and can be as simple as a linear model or more sophisticated machine learning models. Therefore, each created energy consumption model is normalized against the weather variation but not to exogenous factors that are discussed in the previous section. Thus, if a simple aggregation (e.g., sum of savings) is performed to evaluate the impact of the EE program, the resulting estimate does not account for non-routine events, free-ridership, or spillover.

Advanced M&V software tools developers are exploring the inclusion of comparison group approaches in their methods. In the following we will discuss the EM&V techniques used by EnergySavvy and Recurve, which are two software providers that have been prominent in development and discourse on residential M&V methods.

EnergySavvy's M&V 2.0 software tool is based on an approach described in the Uniform Methods Project (UMP) chapter 8 (Agnew and Goldberg 2017), which is called a two stage approach. This approach can be described as follow:

- Stage 1 Individual customer analysis: for each customer within the participant and the comparison group estimate the difference in the energy consumption (i.e., savings) between the pre and post period using regression models.
- Stage 2 Cross-sectional Analysis: combine the estimates obtained in stage 1 to estimate the aggregate EE program impact.

In Addition, EnergySavvy's tool intends to mitigate the effect of bias due to exogenous factors by building aggregate models of the bias and variance using data from the non-participants group (EnergySavvy 2017):

- The bias model is used to adjust the savings estimated for each EE program participant to "account for correlated changes in usage observed in the general population".
- The variance model is used to produce an empirical estimate of uncertainty in the predicted energy consumption, which is used to obtain the prediction interval of the aggregated savings.

These models (i.e., bias and variance) are built using a machine learning algorithm (i.e., random forests) and several features as independent variables. The number of features will depend on the needs of each program and available data. For each EE program participant, the bias and the variance are estimated using the relevant features of the customers. The estimated bias is subtracted from the participant's predicted energy consumption and thus a bias-adjusted savings estimate is produced. While EnergySavvy describes the bias and variance models there is currently no independent documentation of their effectiveness.

Energy Trust of Oregon and Recurve have recently conducted a study to evaluate several matching methods of comparison group (Shaban and Young 2018). They have considered 5 different approaches:

- Random sampling: methods that involve selecting a random sample of non-participants that was equal to 5 times the number of customers in the participant group.
- Future participants: Using customers that will participate in the EE program in the future as a comparison group.
- Stratified sampling: the strata were defined by splitting the customers into deciles of annual energy consumption. Within each stratum a random sampling was performed. The number of customers in the comparison group within each stratum was equal to 5 times the number of customers in the participant group that were part of the same stratum.
- Annual consumption matching: involves finding for each participant 5 "nearest neighbors" based only on the annual energy consumption.
- Monthly consumption matching: for each customer of the participant group 5 customers from the non-participant group are selected as a match using the Euclidean distance metric.

The considered methods were assessed using six different equivalence metrics. Three of these metrics were visual (i.e., annual consumption histogram, annual consumption Q-Q plot and plot of monthly energy consumption during the baseline), whereas the other three are quantitative (i.e., P-value from the t-test of annual energy consumption for the participants and comparison group, P-value from the t-test of monthly energy consumption for the two groups and the Kolmogorov-Smirnov test for monthly energy consumption for the two groups). The dataset used in the Energy Trust of Oregon study consisted of approximately 600 EE program participants and a pool of ~200,000 non-participants was used to determine the comparison group.

The main finding of the Energy Trust of Oregon and Recurve study were that the differences in the savings estimates of different matching methods were not always statistically significant. Thus, the authors of this study were not able to define just one best matching method. However, they provided the following recommendation to define the comparison group:

- "Use up to three methods of comparison group identification depending on data availability". The three methods that were recommended by the study were the matching on monthly consumption using the Euclidean distance, stratified sampling of future participant groups and stratified of past participant groups.
- Use of several metrics to evaluate the similarity between the participant and the comparison groups.

- Select the comparison group customers from the same zip code as the corresponding participants.

Recurve is planning to include the recommended matching methods into their software tool. This new feature will follow a two-stage approach described in the Uniform Methods Project (UMP) chapter 8 (Agnew and Goldberg 2017). At the time of writing this memo it was still under development.

4. Conclusions

There is increasing attention in using automated and standardized tools to evaluate EE programs, due to increased availability of advanced metering that can provide higher granularity data (e.g., daily, hourly) and recent advances in the development of analytical methods. The combination of advanced M&V methods with EM&V best practice use of comparison groups could be a promising step toward getting a timely, accurate assessment of program savings. The standard practice of advanced M&V does not consider the effect of exogenous factors and thus only provides the gross savings estimates. If advanced M&V tool vendors add QEM, and more specifically the usage of comparison group, this could offer a possible solution to reduce the bias produced by these exogenous factors. Although these techniques will probably not completely eliminate the bias they are the most promising solution if the RCT approach is not feasible.

Advanced M&V for residential applications offer two significant benefits to utilities. Firstly, they provide an opportunity to assess program savings in a timely manner using automated analytics, rather than waiting for formal evaluation results several years after program end. Secondly, advanced M&V tools open up opportunities to develop innovative pay-for-performance program designs. As illustrated in this memo, current evaluation practices include a variety of methods, all of which include some level of compromise between accuracy/bias and the practicalities of time/budget/data availability. Likewise, advanced M&V tools offer varying methods, and there is a lack of independent literature verifying their performance and accuracy, hence the importance of public demonstration projects employing these advanced M&V tools.

References

Granderson, J., S. Touzani, C. Custodio, M. D. Sohn, D. Jump, and S. Fernandes. 2016. “Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings.” Applied Energy, 173, pp.296–308.

Granderson, J., Touzani, S., Fernandes, S. and Taylor, C., 2017. “Application of automated measurement and verification to utility energy efficiency program data”. Energy and Buildings, 142, pp.191-199.

Granderson, J. and Fernandes, S., 2017. The state of advanced measurement and verification technology and industry application. The Electricity Journal, 30(8), pp.8-16.

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

Violette, D. M.; Rathbun, P. (2017). Chapter 21: Estimating Net Savings – Common Practices, The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68578. http://www.nrel.gov/docs/fy17osti/68578.pdf

SEE Action (2012). Energy Efficiency Program Impact Evaluation Guide. Prepared by Evaluation, Measurement, and Verification Working Group. https://www4.eere.energy.gov/seeaction/system/files/documents/emv_ee_program_impact_guide_0.pdf

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

EnergySavvy (2012). “M&V 2.0 Detailed Methodology as applied to the estimation of savings from energy efficiency programs”.

Shaban, H. Young, M. (2018). “Comparison Group Identification for Impact Evaluation”; Energy Trust of Oregon. https://www.energytrust.org/wp-content/uploads/2018/11/OpenEE-Technical-Report-Comparison-group-identification-methods-FINAL-wSR.pdf