# Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol

## The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

**This version supersedes the version originally published in April 2013. The content in this version has been updated.**

Ken Agnew and Mimi Goldberg
*DNV GL*
*Madison, Wisconsin*

NREL Technical Monitor: Charles Kurnik

**NREL is a national laboratory of the U.S. Department of Energy**
**Office of Energy Efficiency & Renewable Energy**
**Operated by the Alliance for Sustainable Energy, LLC**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

**Subcontract Report**
NREL/SR-7A40-68564
November 2017

Contract No. DE-AC36-08GO28308

NREL
NATIONAL RENEWABLE ENERGY LABORATORY

# Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol

## The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

**This version supersedes the version originally published in April 2013. The content in this version has been updated.**

Ken Agnew and Mimi Goldberg
*DNV GL*
*Madison, Wisconsin*

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

**NREL is a national laboratory of the U.S. Department of Energy**
**Office of Energy Efficiency & Renewable Energy**
**Operated by the Alliance for Sustainable Energy, LLC**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

**Subcontract Report**
NREL/SR-7A40-68564
November 2017

Contract No. DE-AC36-08GO28308

# Disclaimer

These methods, processes, or best practices ("Practices") are provided by the National Renewable Energy Laboratory ("NREL"), which is operated by the Alliance for Sustainable Energy LLC ("Alliance") for the U.S. Department of Energy (the "DOE").

It is recognized that disclosure of these Practices is provided under the following conditions and warnings: (1) these Practices have been prepared for reference purposes only; (2) these Practices consist of or are based on estimates or assumptions made on a best-efforts basis, based upon present expectations; and (3) these Practices were prepared with existing information and are subject to change without notice.

The user understands that DOE/NREL/ALLIANCE are not obligated to provide the user with any support, consulting, training or assistance of any kind with regard to the use of the Practices or to provide the user with any updates, revisions or new versions thereof. DOE, NREL, and ALLIANCE do not guarantee or endorse any results generated by use of the Practices, and user is entirely responsible for the results and any reliance on the results or the Practices in general.

USER AGREES TO INDEMNIFY DOE/NREL/ALLIANCE AND ITS SUBSIDIARIES, AFFILIATES, OFFICERS, AGENTS, AND EMPLOYEES AGAINST ANY CLAIM OR DEMAND, INCLUDING REASONABLE ATTORNEYS' FEES, RELATED TO USER'S USE OF THE PRACTICES. THE PRACTICES ARE PROVIDED BY DOE/NREL/ALLIANCE "AS IS," AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL DOE/NREL/ALLIANCE BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER, INCLUDING BUT NOT LIMITED TO CLAIMS ASSOCIATED WITH THE LOSS OF PROFITS, THAT MAY RESULT FROM AN ACTION IN CONTRACT, NEGLIGENCE OR OTHER TORTIOUS CLAIM THAT ARISES OUT OF OR IN CONNECTION WITH THE ACCESS, USE OR PERFORMANCE OF THE PRACTICES.

# Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

# Acknowledgements

The authors of this chapter wish to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Jessica Granderson of Lawrence Berkeley National Laboratory
- M. Sami Khawaja of Cadmus
- Frank Stern, Carly Olig, and Pace Goodman of Navigant
- Tim Guiterman of Energy Savvy
- Katherine Randazzo and Seth Wayland of Opinion Dynamics
- Rick Ridge of Ridge and Associates
- Kevin Warren of Warren Energy.


# Suggested Citation

Agnew, K.; Goldberg, M. (2017). *Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol, The Uniform Methods Project: 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

# Acronyms

| | |
|---|---|
| AMI | advanced metering infrastructure |
| CO | cooling only |
| DOE | U.S. Department of Energy |
| HC | heating-cooling |
| HER | Home Energy Report |
| HO | heating only |
| IPMVP | International Performance Measurement and Verification Protocol |
| IV | instrumental variables |
| LATE | local average treatment effect |
| NAC | normalized annual consumption |
| NOAA | National Oceanic and Atmospheric Administration |
| NREL | National Renewable Energy Laboratory |
| RCT | randomized control trial |
| RED | Random Encouragement Design |
| SAE | Statistically Adjusted Engineering |
| TMY | typical meteorological year |
| UMP | Uniform Methods Project |

# Protocol Updates

The original version of this protocol was published in April 2013.

This chapter has been updated to incorporate the following revisions:

- Clarified scope for the chapter. This chapter can be used for non-whole-house programs but is not the only approach for whole-house. Clarified application of discussion to daily data without full exploration of options with daily data.
- Expanded allowable modeling options. Included pooled with comparison group, randomized encouragement design, instrumental variables and inverse Mills ratio. Re-worked discussion of and recommendations related to participant-only pooled approach.
- Clarified language (no fundamental changes) around comparison group and net savings.

# Table of Contents

# List of Tables

# 1 Measure Description

Whole-building retrofits involve the installation of multiple measures. Whole-building retrofit programs take many forms. With a focus on overall building performance, these programs usually begin with an energy audit to identify cost-effective energy efficiency measures for the home. Measures are then installed, either at no cost to the homeowner or partially paid for by rebates and/or financing.

The methods described here may also be applied to evaluation of single-measure retrofit programs. Related methods exist for replace-on-failure programs and for new construction, but are not the subject of this chapter.[1]

[1] As discussed in the section "Considering Resource Constraints" of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

# 2 Application Conditions of Protocol

The estimation of the total savings from a multi-measure project requires a comprehensive method for capturing the combined effect of the installed measures. The general method recommended for this type of program is the analysis of consumption data from utility billing records. This method has traditionally been referred to as a billing analysis, and is referred to in this chapter as consumption data analysis.

Unlike the evaluation methods described in most of the other measure-specific chapters of the Uniform Methods Project (UMP), the whole-building analysis methods described in this chapter are designed to provide savings for a program or program segment and do not necessarily produce savings for each participating building. These program-level methods apply only for populations of relatively homogenous buildings. Program-level consumption data analysis as described in this chapter is most commonly applied to residential buildings.

At the individual building level, these methods are consistent with the general approach of International Performance Measurement and Verification Protocol (IPMVP) Option C, Whole Facility. Option C is designed in part to address evaluation conditions that occur with a whole-building retrofit program. However, the IPMVP is designed for individual building analysis, and Option C includes explicit adjustment for non-routine changes, that is, for changes unrelated to the measures of interest that also affect energy consumption. By contrast, the consumption data analysis methods described in this chapter use analysis across multiple buildings to control for non-program-related changes. For a whole-premise program with very heterogeneous participants, such as a large commercial buildings program, the methods described in this chapter are not well suited.

The consumption data analysis approach has strengths and limitations that render it more appropriate to certain types of whole-building program evaluations than to others. This chapter describes how a consumption data analysis can be an effective evaluation technique for whole-building retrofit programs, and it addresses both how and when consumption data analysis should be used.

The evaluation methods noted in this chapter for monthly consumption data are applicable when all of the following are true:

- The whole-building savings from the combination of measures supported by the program are expected to be of a magnitude that will produce statistically significant[2] results given:
    - The natural variation in the consumption data
    - The natural variation in the savings
    - The size of the evaluation sample.

[2] The required level of statistical accuracy may vary from across evaluations. In addition, statistically significant difference from zero is a substantially lower bar than the frequently proposed 90/10 relative precision. Relative precision depends on the magnitude of the savings estimate, the number of participants and, depending on approach, the granularity of the data. Billing analysis results that have 90/50 relative precision are common and may provide acceptable results for the purposes of a program evaluation.

- The baseline for determining savings is the condition of the participating building before the retrofits were made, rather than the standard energy efficiency of the new equipment.
- There is sufficient consumption data available—in the form of monthly or two-month utility billing records—for the participants.
- Consumption data with which to create a comparison group are available for the same timeframe as for the participants for one or more of the following groups:
    - Previous participants—those who took part in the program before the timeframe of the current evaluation
    - Subsequent participants or those who are on a list for future participation in the program
    - Nonparticipants who do not fit either of the first two definitions who are chosen at random or through methods discussed in Section 3.

The evaluation methods described in this protocol are also useful for single-measure programs when all of the requirements listed above are met. Also, note that UMP *Chapter 5: Residential Furnaces and Boilers Evaluation Protocol* uses a consumption data analysis result and addresses the "standard efficiency" baseline issue described in the second bullet above.

## 2.1 Protocol Applicability to Interval Consumption Data

The methods discussed here are presented as applicable to monthly consumption data. Advanced metering infrastructure (AMI) interval data are now available from many billing systems. These data are commonly available at the hourly or even 15-minute level. From the perspective of billing analysis evaluation, such data, especially when worked with at the daily level, are a finer-grained form of the same basic data. The monthly methods discussed here use, as a dependent variable, average daily consumption as developed from the monthly data. The models will also work when applied to actual daily data. This step is conceptually simple and considerably increases the number of data points available for a single building in a year.

These finer-grained data that are available move consumption data modeling into a wider realm that is beyond the scope of this protocol as initially defined. In addition to likely improvements from more nuanced models that leverage the additional degrees of freedom and more direct relationship between consumption and weather, there are challenges to using daily data. One concern is the increased serial correlation in the modeling process with the more granular data. Hourly data open up a still wider array of issues as the diurnal patterns combined with the unique thermal dynamics of each building demand more complex statistical treatment. A great deal of exploratory work has been done, primarily in the commercial space where interval data have been available for longer. A future protocol will address these issues as methods are consolidated.

# 3 Savings Framework

Energy consumption data, with their wide availability and their explicit tracking of consumption changes over time, appear to offer a straightforward approach to measuring energy savings. The potential for the change in consumption between two periods to be linked to a program implementation or treatment is compelling. A primary challenge is isolating the effect of the treatment from other sources of change over the same period.

## 3.1 Components of Change in Consumption

An observed change in consumption between pre- and post-installation periods includes the effect of the whole-building intervention itself, along with the effects of other factors unrelated to the program that may occur in the same timeframe. These effects could include changes in occupancy, physical changes to structure, behavioral changes, weather, etc. Without special attention, these non-program effects may be conflated with program effects leading to incorrect estimates of program effects or savings. This chapter focuses on techniques that attempt to address these concerns including regression techniques, comparison groups, and regression techniques with comparison groups.

### *3.1.1 Savings Components Captured*

The participants' change in consumption includes both direct and indirect effects of the program, in addition to the non-program effects. The program-related effects include the following:

- The **direct effect** of the program measures on the affected systems. For example, replacing an existing light bulb with a more efficient light bulb, while using the lights for the same amount of time, reduces electricity used for lighting.
- **Physical interactive effects** between the directly affected system and other systems in the premise. For example, reduced energy use for lighting also reduces the need for cooling in the summer and increases the need for heating in the winter.
- **Take-back or rebound effects,** where a system is used more *because* it has been made more efficient. For example, the household might pay less attention to turning lights off because the cost of extra lighting use is lower.
- **Participant same-year spillover effects**, where participants install additional energy-savings equipment *because* of the program but *outside* of it, within the post-installation study period. For example, a positive experience with program-provided efficient lighting might lead to installing efficient lighting in other places, or to taking on additional non-lighting efficiency upgrades learned about or encouraged by the program.

The billing analysis does not separate these effects. All are included in the savings estimate captured by the analysis. To the extent takeback or spillover are delayed responses to the measure installation, the measured savings might include only a partial year of these effects. Participant spillover that occurs beyond the timeframe of the post-installation period studied is not captured at all.

Conversely, to the extent that nonparticipants have undertaken measures as a result of the program but outside it, the billing analysis does not capture that nonparticipant spillover as part of the savings. To the contrary, nonparticipant spillover that occurs within the same timeframe as

the participant installations may reduce the estimated savings due to the program, since the effect will be treated as part of the exogenous change as captured by a comparison group.

The timeframe of spillover is an important consideration in understanding what the billing analysis results provide. Nonparticipant spillover from past programs that affect the program-year actions of both current-year participants and current-year nonparticipants can be seen as part of the current-year market condition. The billing analysis is capturing savings relative to that current market condition. Spillover or market effects from prior years need to be addressed by different methods than those described here. Spillover to future years outside the analysis period also has no effect on the billing analysis and is not addressed by it.

Contemporaneous nonparticipant spillover due to the current year program and occurring within the same time frame may count against the program in the billing analysis, to some extent. How much the estimated savings is biased downward by the nonparticipant spillover depends in part on the timing of the spillover within the study period, relative to the timing of participation. The bias will be on the order of twice the average contemporaneous nonparticipant spillover savings per household, or less depending on timing.

### 3.1.2 Free-Ridership as a Component of Savings

The prior two sections highlight the challenge of isolating the program effects of interest (savings) and understanding what components make up those savings. A final aspect of savings that will be addressed throughout the remainder of this chapter is the presence of free-ridership or the difference between net and gross savings. There is a separate UMP chapter that addresses the challenge of free-ridership across evaluation generally (Violette 2017); however, because the choice of consumption data analysis approach has implications for the assessment of free-ridership, it is necessary also to address the subject here.

Chapter 21 of the UMP defines free-ridership as "the program savings attributable to free-riders (program participants who would have implemented a program measure or practice in the absence of the program)." That chapter also provides the following definitions of gross and net savings:

- **Net savings:** The difference in energy consumption *with the program in place* versus what consumption would have been *without the program in place*.
- **Gross 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*.

The consumption data analysis approach has implications for whether savings estimates are gross, net or somewhere between.

## 3.2 Comparison Group Specification

Comparison groups play an important role facilitating the isolation of program effects across a range of disciplines. A comparison groups offers a proxy counterfactual against which an effect can be estimated. In some disciplines a comparison group is used to support an estimate of treatment effect even when data from before and after treatment is not available. The combination of comparison groups with pre- and post-treatment data have the potential to

support more robust results than either approach on its own. While there is discussion in Section 4 of approaches that forego comparison groups completely, most consumption data analysis includes some form of implicit or explicit comparison group.

A comparison group consisting of general nonparticipants drawn from the eligible population will control to some extent for factors in the market affecting all customers, such as changes in the economy, prices of energy and of energy-using equipment, or weather.[3] However, these exogenous changes affect different types of customers differently. These different responses to the external factors stem from both the different physical structure and equipment, and the different customer behavioral and decision characteristics. Thus, it's important for the comparison group to match the participants as well as possible in terms of both physical and behavioral/decision-making characteristics.

There are several possible ways to develop a comparison group that is well matched to the participant group in terms of these two dimensions:

1. Starting from a pool of similar customers, randomly assign some customers to receive the whole-building treatment and the others not to. This randomized control trial (RCT) approach is ideal for unbiased measurement of the treatment effect, but for most programs is unrealistic and inconsistent with the program delivery mechanism and theory.
2. Use future program participants as a comparison group for those who participated in the current program year. If the program and participant mix are stable over time, future participants will be similar to current participants, apart from the participation itself.
3. Use past program participants as a comparison group for those who participated in the current program year. The concept is similar to the use of the future participants as a comparison group, as explained further below.
4. Use a set of nonparticipants chosen to match the participants on observable characteristics. Matching characteristics can include consumption in earlier periods, demographic information known from customer records, geography, or explicit average census variables determined from geography. Self-selection bias is still a concern with matched comparison groups, as described below.

### 3.2.1 Randomized Controlled Trial Control Group

The optimal evaluation scenario for a consumption data analysis is an RCT experimental design. This is the standard approach used across the experimental sciences to (1) isolate treatment (program) effects and (2) establish a causal link between the treatment and the effect.

While a control group constructed by random assignment via an RCT is the "gold standard" for a comparison group, this approach is not practical for most programs. For an RCT, a pool of eligible participants is randomly assigned to one of two groups before the program engagement. This random assignment process assures that the two groups—treatment and control—are similar on average in every respect except for the offer of program treatment. In this context, eligibility

[3] While weather-related change is a form of exogenous change, it is controlled for as well as possible by the weather-normalization process of the Stage 1 models. To the extent that weather is incompletely normalized, the inclusion of a comparison group that has also been weather-normalized will control for any remaining uncontrolled for weather effects.

requirements are defined by those running the program and setting up the RCT (top quartile of consumption, etc.). RCT results will be applicable for the treatment group and other customers that meet eligibility requirements who receive the same treatment (external validity).

The basic analysis structure commonly applied with an RCT is a "difference of differences" calculation. The program-related change is estimated as the difference between the treatment group pre-post difference and the control group pre-post difference.

- For the treatment group, the pre-post difference includes the program-related change plus exogenous change.
- For the control group, the pre-post difference includes only exogenous change.

Because of the random assignment, the average exogenous change is expected to be the same for the treatment and control groups. The control group estimate of exogenous change is used to adjust the treatment group, removing or controlling for that exogenous change. The adjustment is additive and may be positive or negative depending on the direction of the exogenous trend. The final result is an estimate of the treatment group's program-related change.

In the context of energy efficiency programs, true RCT is rare outside of certain types of behavioral programs.[4] The approach, however, provides a good illustration of the ideal characteristics of a comparison group. In particular, the RCT scenario provides an example where the resulting savings estimates are net savings with the effects of free-ridership controlled for. In non-RCT design, the results are generally not net savings. This issue is discussed in Section 3.2.2.

A Random Encouragement Design (RED) approach offers a more flexible approach to incorporating random assignment and addressing challenges related to net savings. We discuss this approach in Section 3.5 after the full range of challenges has been described.

### 3.2.2 Non-Randomized Comparison Group Development

Where program delivery is not designed as an RCT or other random assignment design, a comparison group is developed after the fact in a quasi-experimental design framework. For that design framework, the term "comparison group" denotes groups that are not randomly assigned but still function similarly to experimental control groups.

The comparison group, which is designed to be as similar as possible to the treatment group during the pre-evaluation period, can be matched to the treatment group using a variety of known characteristics such as geography and pre-program consumption levels.[5] In this context, the

[4] There are multiple reasons why RCT has not been more widely employed. RCT requires denying or delaying participation to a subset of the eligible, willing population and could involve forcing services on people who either do not want them or may not use them. Regulators generally do not support providing tangible services to some customers and not others, outside of limited pilot situations. RCT works for behavior/information programs because there's no forced interference with the premise (recipients can opt out with the utility or effectively opt out by ignoring the reports), and there's no tangible service restricted to only the treatment group.

[5] Since the original writing of this protocol, matched comparison approaches have gained wide acceptance for certain kinds of programs where savings are expected to be small and an RCT control group is not available. Opt-in behavior programs are an example of this kind of program. The limitations of the approach are recognized but no

eligible population from which a comparison group should be drawn follows the requirements for program participation, as much as possible. For example, if participants are required to be single-family dual-fuel households from the utility territory, then the eligible population for a comparison group would start from that definition if that information is available for the general population. As with the true experimental control group, the comparison group is intended to exhibit all of the exogenous, non-program-related effects due to the economy and other factors affecting energy consumption. Thus, the comparison group provides an estimate of exogenous change to use in adjusting participant pre-post impacts.

Unfortunately, matching a comparison group to the treatment group on known characteristics does not produce a true control group. Most importantly, post-hoc matching does not address the issue of self-selection. By the very decision to self-select into a program, the members of the treatment group are different from those of any comparison group that can be constructed post-hoc from nonparticipants.

In theory, many important characteristics can be controlled for in matching or screening to construct the comparison group; however, in reality, the available characteristic data on the customer population is relatively sparse. Also, some important characteristics—such as environmental attitudes—are effectively unobservable. The result is a potential bias that cannot be quantified.

In the context of an energy efficiency program evaluation, the issue of self-selection is complicated by the added dimension of free-ridership. 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. This characteristic, being a "natural adopter," is unobservable for both participants and nonparticipants. Even for customers who match closely on observable characteristics, those who would adopt on their own are more likely to join a program than those who would not. As a result, self-selection affects the ability to obtain an unbiased estimate of savings, and it affects whether that estimate of savings is best considered gross, net, or something in between.

## 3.3 Practical Match Comparison Group Development

In some cases, it is not practical to use past or future participants as a comparison group, nor to conduct a pooled[6] consumption data analysis with participation staggered across a year or more. This tends to be the situation when one or more of these conditions are present:

- The program has not been stable over previous and subsequent years.
- The program has not had consistent data-tracking over a sufficient length of time.
- The program participation effects extend over a long time after the tracked participation date, e.g., multiple installation dates, or delayed effects as from a behavioral intervention.

alternative exists. As a result of this work, matching methods such as propensity score matching and minimum distance algorithms have seen wide usage. The specifics of these approaches are beyond the scope of this protocol.

[6] Through this chapter we use the term "pooled" to refer to time-series, cross-sectional data and models.

- The program roll-out results in all participation occurring during only a few months of the year. In such a case, the pooled method will not be useful unless multiple years of participation can be included in the model.

In these cases, either a two-stage or pooled model using a matched nonparticipant comparison group is recommended. One condition for using the general eligible nonparticipant population as a comparison group is that the characteristics of the nonparticipants should be generally similar to those of the participants. Typically, this is not the case. Thus, when participants are different—on the whole—from nonparticipants, a matched group of eligible nonparticipants provides a better comparison group to control for non-program factors among similar premises. However, a matched nonparticipant group is still subject to the same kinds of biases related to naturally occurring savings, self-selection, and spillover, as described above for the general eligible nonparticipant population.

One type of matching stratifies the participants and the comparison pool by observable characteristics, and then randomly selects comparison cases for each stratification cell, proportional to the number in the participant group. Thus, once the matching variables and their ranges or levels are decided, the process is (1) determining the proportion of the participant population in each cell and (2) selecting a nonparticipant sample with the corresponding proportions, from those customers who satisfy the basic eligibility requirements. The following matching factors may be used, depending on their availability:

- Consumption level or other size measure
- Demographics, especially income and education
- Dwelling unit type
- Geography (ZIP code, if feasible)
- Energy end uses.

Another form of matching assigns one (or a specified multiple) specific matched comparison customer(s) to each participant. Propensity score matching and minimum difference algorithms can be used to develop such matching at the customer level across the population or within strata. A variety of approaches for matching are available and new approaches are being tested (Machine learning (e.g., random forests), etc.).

## 3.4 Self-Selection and Free-Ridership

Whenever a comparison group is selected from customers who were eligible to join the program but chose not to, the potential for self-selection bias is a concern. That is, customers who chose to participate in the program (at a particular time) may have systematic differences from those who did not, resulting in systematic differences in their (changes in) energy consumption, apart from the effects of the program itself. These systematic differences can lead to bias in the savings estimate. While the comparison group construction can control from some of these differences, there are some key differences it can't control for.

A comparison group of eligible nonparticipants controls, in part, for general factors affecting the market, but the general nonparticipant group may respond differently to these general factors than the participants would have without the program.

Matching on observable characteristics, or explicitly including characteristics variables in a regression model, improves the ability of the comparison group to control for the exogenous factors as they affect the participants. However, such matching can't control for the largely unobservable decision-making factors that led the participants to join when they did and the comparison customers not to.

The interaction between self-selection and free-ridership is best illustrated with an example. A true control group is similar to the treatment group with respect to natural levels of energy efficiency activity. For example, if 5% of a population would have installed an energy-efficient furnace without rebate assistance, then the same percentage of both the treatment and control group populations will exhibit this behavior. In the treatment group, some or all of this 5% will participate in the program. By definition, this set of participants consists of free-riders.

In the RCT scenario, the control group does not have access to the program. The naturally occurring savings generated by the 5% natural adopters of the measure in the control group is part of the pre-post non-program exogenous change. The savings from this 5% of the control group that are natural adopters of the measure will equal (on average) the savings for the 5% natural adopters in the treatment group. This natural-adoption portion of treatment-group savings will thus be cancelled out by the corresponding naturally occurring adoption in the control group in the difference of differences calculation. That is, in a true RCT design, naturally occurring energy efficiency savings—and, in the process, free-ridership—are fully removed from the estimate of program-related savings. The result is a "net" estimate of savings; that is, program savings net of free-ridership.

By contrast, an evaluation using a post-hoc comparison group will not generally produce a net savings result. In a non-RCT program scenario, the 5% of households naturally inclined toward the measure adoption all have the option to opt into the programs. Unlike the even allocation across treatment and control groups in the RCT scenario, the allocation of the non-RCT scenario depends on the rate of strategic behavior by the adoption-inclined population. Customers and contractors inclined toward adopting the measure have little reason not to take advantage of the program. This inclination is likely to lead to higher proportion of natural adopters in the participant population, as compared to the general incidence in the population. This differential proportion of natural adopters then affects in multiple ways the level of savings and free-ridership that will be measured by the consumption data analysis.

- First, the participant group includes a higher proportion of natural energy efficiency adopters than would a randomly assigned treatment group (or the general eligible population), due to self-selection into the program. These natural adopter households that strategically opt into the program increase the free-ridership rate among program participants beyond the natural proportion of natural adopters in the eligible population.
- Second, it follows from this that any comparison group developed after the fact from those who chose not to participate will tend to have a lower percentage of natural energy efficiency adopters than would a randomly assigned control group. To return to the

scenario where 5% of the overall population are natural energy efficient furnace adopters, the reduced presence of natural adopters in the comparison group population (<5%) will not negate the self-selected and oversized presence of natural adopters in the participant group (>5% up to 100%).

- Finally, the concerns regarding self-selection beyond the issue of natural adoption are still present. Related to their natural inclination to adopt energy efficiency, the program participants may exhibit different energy-consumption patterns, and different consumption change patterns than the general population. Matching algorithms can help to match the observable characteristics of the comparison group to the participant group. However, the matching inherently cannot match on non-program-induced consumption changes, which are unobservable. To the extent that participation is related to such changes, matching approaches will not fully address self-selection and any associated biases.

These are the key factors that make it impossible for the matched comparison to fully reflect the non-program changes among the participants. As a result, when comparison group change is netted out of the participant change, the netting will control for some but not all of the naturally occurring measure implementation, leaving an unknown amount of free-ridership in the final savings estimate. The resulting estimate is thus somewhere in between net and gross savings.

In the extreme, all households that naturally install an energy-efficient furnace will purchase through the program, leaving no natural energy efficiency purchasing in the non-program population from which the comparison group is constructed. Under this extreme scenario, the comparison group would only provide an estimate of exogenous change apart from natural measure adoption, and would not control for any natural energy efficiency activity. This savings estimate would retain all of the free-rider savings and, thus, would best be classified as a gross savings estimate.

The general recommendations in this whole-building retrofit protocol address these issues by constructing comparison groups that are composed of customers who have opted into the same program in a recent year—or will participate in the near future (pipeline). This approach avoids concerns related to self-selection bias in two ways. Because they have participated or will participate in the same program, they are similar to the participants being evaluated with respect to energy consumption characteristics.[7]

Just as importantly, because they have just participated (or soon will participate) in the program, these previous and future participants are unlikely to install the program measures on their own during their non-participating years.[8] As a result, a comparison group created from previous and

[7] See Randazzo et al. 2017 for an alternative perspective.

[8] If some program-eligible measures are installed without support of the program during the period of time used for the analysis, then the effects of those outside-program installations would be included with the other exogenous change captured by the comparison group. The participants under evaluation would be expected to install outside the program at a similar rate. Depending on the timing of the outside-program installations relative to the timing of participation, some bias can be introduced in either direction. However, if the outside-program installations are timed similarly for current and future participants, and are spread over something like two or more years prior to participation, the future participants will correctly control for current participant outside installations and bias will be minimal.

future participants may be as similar to current-year participants (apart from the program effect itself) as is possible outside of a random assignment design. Thus, the use of such a comparison group is likely to produce a gross estimate of savings that is less biased with respect to self-selection.

## 3.5 Random Encouragement Design

A Random Encouragement Design (RED) uses random assignment but is more practical to implement for many programs than an RCT. Under the RED, the eligible pool of customers is randomly assigned either to receive supplemental encouragement to participate in the program or not to receive that encouragement. Supplemental encouragement may consist of higher incentive levels, or expanded outreach.

With the RED, a basic difference-of-differences analysis subtracts the average change in consumption among customers who received the supplemental encouragement from the average change among customers who did not receive supplemental encouragement. The averages are calculated across all the customers in each group, not just program participants. The difference of differences is the average change in consumption associated with incremental encouragement. Dividing this difference by the difference in participation rate between the two groups gives the average change in consumption per incremental participant due to encouragement. This incremental change per incremental participant is known as the local average treatment effect (LATE).

A variant of the difference of difference analysis uses a regression approach with instrumental variables (IV), as described in Section 4.3.3. The simplest form of this regression is equivalent to the difference of differences LATE calculation, and provides the same result. A more informed version uses additional explanatory variables.

Regardless of whether difference of differences or basic IV regression is used, the RED produces net savings for the program of interest only under restricted circumstances. The RED does produce incremental net savings per incrementally encouraged participant. However, this incremental savings per incremental participant is not the same as the savings per participant in the base (no-encouragement) program, and in fact may be very different from the base program's net savings. In particular, we anticipate that customers who participate only with supplemental encouragement are less likely to be free-riders than those who participate in the base program. Thus, the RED with basic IV analysis is likely to overstate net savings per participant for the base program. This approach is likely to give an unbiased estimate of net savings for the base program only if:

1. Free-ridership is minimal—that is, net and gross savings are the same
2. There is no relationship between how much energy a customer will save by participating and their inclination to participate (Goldberg et al. 2017).

Nevertheless, obtaining an estimate that can be regarded as a likely upper bound on net savings may itself be useful.

# 4 Savings Estimation

## 4.1 Recommendations by Program Characteristics

The consumption data analysis specification and interpretation depend on both the program structure and the corresponding comparison group specification. For a variety of program characteristics, Table 1 shows how the comparison group can be specified and how the resulting savings should be interpreted. Note that some program structures are best for determining net savings, while others are best for determining gross savings. The "consumption data analysis form" column refers to two-stage and pooled modeling approaches which are discussed at length in sections 4.3 and 4.4, respectively.

**Table 1. Program Characteristics, Comparison Group Specifications, and Consumption Data Analysis Structure and Interpretation**

| Randomized Controlled Trial? | Stable Population? | Comparison group | 2-Stage and Pooled with Comparison Group | Gross or Net Savings | Unknown Biases |
|---|---|---|---|---|---|
| Yes | N/A | Randomly selected control group | Yes | Net | Spillover from T to C, if it exists |
| No | Yes | Prior and/or future participants | Yes | Gross | Time-varying Characteristics |
| No | No | Matched comparison group | Yes | Likely between gross and net | Time-varying Characteristics, Self-selection unaccounted for by matching and same-period NP spillover |
| No | No | General eligible nonparticipants | Yes, With additional characteristics in the 2nd stage or pooled regression | Likely between gross and net | Time-varying Characteristics, Self-selection unaccounted for by regression and same-period NP spillover |

Table 1 provides a rough order of preference for analysis form as program conditions become less ideal. Importantly, each approach has strengths and weaknesses that, in specific evaluation scenarios, might justify choosing an approach from lower in the table.

1. **Randomized controlled trial experimental design.** The RCT scenario is unique in that consumption data analysis form will not affect the unbiasedness of the treatment effects. Pooled models will generally provide additional power and specifically, lagged dependent variable models have become a standard approach in the Home Energy Report (HER) literature. These models are discussed in UMP Chapter 21 (Stewart et al. 2017). HER RCT models are almost always designed to measure actual-weather savings, so this modeling approach also avoids distinctions between two-stage and pooled with respect to weather-normalization.

2. **Not randomized, stable program and target population over multiple years.** Table 1 recommends either a two-stage or pooled model with the comparison group created from prior or future participants. Stability, in this case, refers primarily to changes in eligibility rules or major shifts in the supported measures. Changes in targeting and/or marketing may or may not have similar effects. As discussed, the use of the prior/future participants has the potential to address many of the concerns related to self-selection while delivering an estimate of gross savings.[9] The results from the two-stage and pooled approaches should be similar. The two-stage approach offers the increased flexibility with respect to weather modeling relative to the single, mean weather effect estimated in the pooled model. The pooled approach will provide relatively greater precision.
3. **Not randomized, not stable program.** Table 1 recommends a matched comparison group in either a two-stage or pooled approach. As discussed, the matched comparison group should address self-selection bias with respect to the observable characteristics used for matching but not of the remaining self-selection concerns. The savings estimates from this approach will fall somewhere between net and gross. In general, this makes the match comparison group less desirable than the prior/future comparison group. However, in addition to questions regarding program stability for the prior/future approach, prior/future participants will always be relatively less numerous than the eligible matching group. It may be justifiable to use a matched comparison group in place of or in addition to the prior/future participant comparison group. Generally, these results are treated as gross estimates of savings and a separate free-ridership analysis is required (for example, self-reported) to adjust these savings estimates to net savings estimates.
4. **Not randomized, not stable program without matching.** Table 1 offers the final option of a general population comparison group. This approach is similar to the match comparison group approach but with regression variables accounting for differences between the treatment and the more general comparison group. In theory, this approach could be as effective as the matched comparison group. In practice, the data to control for these differences are not readily available. Furthermore, were such variables available, they could also be used either in the matching algorithm or included in the regression with the matched comparison.

## 4.2 Full-Year and Rolling Analysis Using Prior or Future Participants as the Comparison Group

There are two primary ways to structure the analysis with past and future comparison groups: full year and rolling.

### *4.2.1 The Full-Year Specification*

The full-year approach, illustrated in Table 2, compares the energy consumption from the full year ***before*** the current program year to the full year ***after*** the current program year. Thus, the comparison group consists of customers who either (1) participated in the year that ended a year

[9] Low income programs are a good example of a program that can be stable over time. In the case of low income programs, there is limited expectation of natural occurring savings activity so gross savings may be assumed to equal net saving.

before the start of the current program year[10] or (2) participated in the year that began a year after the end of the current program year.

For example, if the program year occurs in calendar year 2011, then savings would be calculated as the change from calendar year 2010 to calendar year 2012, and the comparison group would be participants from calendar year 2009 and/or calendar year 2013.

If the future participants are used, the full-year approach cannot be applied until the group for later years is identified. Few programs have substantial pipelines, so if future participants are to be used, it may be necessary to wait until late enough in 2013 to identify sufficient future participants with 2010 and 2012 data for the evaluation.

**Table 2. Illustration of Analysis Periods for Full-Year Comparison Group, Program Year 2011**

| Group | Participation Timing | Analysis Period 1 (Pre) | Analysis Period 2 (Post) | Expected Change Period 1 to 2 |
|---|---|---|---|---|
| Past Participants | 2009 | Jan 2010 – Dec 2010 | Jan 2012 – Dec 2012 | Non-Program Trend |
| Current-Year Participants | 2011 | Jan 2010 – Dec 2010 | Jan 2012 – Dec 2012 | Program Savings + Non-Program Trend |
| Future Participants | 2013 | Jan 2010 – Dec 2012 | Jan 2012 – Dec 2012 | Non-Program Trend |

### *4.2.2 The Rolling Specification*

Although using the full-year comparison group specification is simple, it requires data from farther back in time. The rolling specification, however, allows data from a more-compressed timeframe to be used, as it uses a rolling pre- and/or post-period across the current program year.

Effectively, for each month of the current program year, this method compares the year ending just before that month with the year that begins after that month. The comparison groups for each month's participation are, therefore, the customers who participated one year before and/or the customers who participated one year later. This structure is illustrated in Table 3 for program year 2011.

[10] Some find it counterintuitive to use past participants for the comparison group because they are no longer similar to pre-program participants by the very fact of their participation. They are, however, assuming a stable program and participation mix, similar in all other ways to post-program participants. The difference-in-differences structure relies on an additive period-to-period change factor that works equally well with past or future participants. Future participants represent how current participants would have changed had they not participated in this year. That is, they capture the effect of all changes other than participation itself. Similarly, past participants represent how current participants would have changed had they already participated prior to this year. Thus, the prior participants also capture the effect of all changes other than participation itself.

**Table 3. Illustration of Analysis Periods for Rolling Comparison Group, Program Year 2011**

| Group | Participation Timing | Analysis Period 1 (Pre) | Analysis Period 2 (Post) | Expected Change Period 1 to 2 |
|---|---|---|---|---|
| Past Participants | Feb 2010 | Mar 2010 – Jan 2011 | Mar 2011 – Feb 2012 | Non-Program Trend |
| | Jun 2010 | Jul 2010 – May 2011 | Jul 2011 – Jun 2012 | Non-Program Trend |
| | Dec 2010 | Jan 2011 – Nov 2011 | Jan 2012 – Dec 2012 | Non-Program Trend |
| Current-Year Participants | Feb 2011 | Mar 2010 – Jan 2011 | Mar 2011 – Feb 2012 | Program Savings + Non-Program Trend |
| | Jun 2011 | Jul 2010 – May 2011 | Jul 2011 – Jun 2012 | Program Savings + Non-Program Trend |
| | Dec 2011 | Jan 2011 – Nov 2011 | Jan 2012 – Dec 2012 | Program Savings + Non-Program Trend |
| Future Participants | Feb 2012 | Mar 2010 – Jan 2011 | Mar 2011 – Feb 2012 | Non-Program Trend |
| | Jun 2012 | Jul 2010 – May 2011 | Jul 2011 – Jun 2012 | Non-Program Trend |
| | Dec 2012 | Jan 2011 – Nov 2011 | Jan 2012 – Dec 2012 | Non-Program Trend |

The comparison group, which captures exogenous change through the evaluation time span, ultimately provides an average of the exogenous change through the 12 months of the current evaluation year. Thus, this group should be selected in such a way that the estimate of exogenous change across the 12 months will be from pre- and post-data periods that are similarly distributed across the evaluation year as the current participants.

If participation rates are stable across the multiple program years being used, the rolling specification will often accomplish a similar distribution over the year without additional effort. However, when using the rolling specification, examine the pattern of participation within each season over the applicable years for each of the two or three groups (current year and past and/or future participants). If the distribution is not similar,[11] then the comparison group should be properly scaled using ***one*** of these methods:

- On a season-by-season basis, sample from the past and/or future comparison groups in proportion to the current year's participation.

[11] This may indicate changes in the program or the program participants that may affect whether this is, in fact, a valid comparison group.

- Re-weight the past and future participants to align with the current-year participants' timing distribution. That is, for a comparison group customer who participated in season s, assign the weight $f_{Ts}/f_{gs}$ where $f_{gs}$ is the proportion of past or future participant group g who participated in seasons and $f_{Ts}$ is the proportion of the current participant group. Then apply these weights in the second-stage analysis.

Generally, for any set of participant sites, the comparison sites need two years of either all pre- or all post-consumption data that cover the year before and after that installation month. This gives the analyst the freedom to create these comparison group pre- and post- data periods using exactly the same distribution as the current year participant dates.

## 4.3 The Two-Stage Approach

### *4.3.1 Stage 1. Individual Premise Analysis*

For each premise in the analysis, whether in the participant or comparison group, do these activities:

1. Fit a premise-specific degree-day regression model (as described in Step 1, below) separately for the pre- and post-periods.
2. For each period (pre- and post-) use the coefficients of the fitted model with normal-year degree days to calculate weather-normalized annual consumption (NAC) for that period.
3. Calculate the difference between the pre- and post-period NAC for the premise.

The site-level modeling approach was originally developed for the Princeton Scorekeeping Method (PRISM™) software (Fels et al. 1995). (The theory regarding the underlying structure is discussed in materials for and articles about the software [Fels 1986].) Stage 1 of the analysis can be conducted using PRISM or other statistical software.[12]

#### *4.3.1.1 Step 1. Fit the Basic Stage 1 Model*

The degree-day regression for each premise and year (pre- or post-) is modeled as:

*Equation 1*

$$E_m = \mu + \beta_H H_m + \beta_C C_m + \varepsilon_m$$

where:

[12] PG&E has supported an effort in California called CalTRACK that is designed to document a set of methods for calculating site-based, weather-normalized, metered energy savings from an existing conditions baseline and applied to single family residential retrofits using data from utility meters, to support various use cases including a residential pay-for-performance pilot. The effort references this UMP chapter, primarily related to Stage 1, site-level modeling. The results of that effort were not finalized at the time of this revision but will offer another source of instruction related to the practical technical methods discussed here. http://www.caltrack.org/

$E_m$ = Average consumption per day during interval *m*
$H_m$ = Specifically, $H_m(\tau_H)$, average daily heating degree days at the base temperature($\tau_H$) during meter read interval m, based on daily average temperatures on those dates
$C_m$ = Specifically, $C_m(\tau_C)$, average daily cooling degree days at the base temperature($\tau_C$) during meter read interval m, based on daily average temperatures on those dates
$\mu$ = Average daily baseload consumption estimated by the regression
$\beta_H, \beta_C$ = Heating and cooling coefficients estimated by the regression
$\varepsilon_m$ = Regression residual.

#### 4.3.1.2 Stage 1 Model Selection

**Fixed Versus Variable Degree-Day Base**

In the simplest form of this model, the degree-day base temperatures $\tau_H$ and $\tau_C$ are each pre-specified for the regression. For each site and time period, only one model is estimated using these fixed, pre-specified degree-day bases.

For ease of processing and of meeting data requirements, the industry standard for many years was to use a fixed 65°F for both heating and cooling degree-day bases. However, actual and normal hourly weather data are easily available now, providing flexibility in the choice of degree-day bases. In general, a degree-day base of 60°F for heating and of 70°F for cooling usually provide better fits than a base of 65°F

The fixed-base approach can provide reliable results if there are only moderate differences between the actual weather used to estimate the models and normal/typical meteorological year (TMY) weather used to construct NAC. When this is the case and data used in the Stage 1 model span all seasons, NAC is relatively stable across a range of degree-day bases. However, the decomposition of consumption into heating, cooling, or baseload coefficients is highly sensitive to the degree-day base. For houses in which the degree-day bases are different from the fixed degree-day bases used, the individual coefficients will be more variable and, potentially, biased as will the combined NAC. As a result, if the separate coefficient estimates will be used for savings calculations or for associated supporting analysis, the fixed degree-day base simplification is not recommended. Similarly, under extreme weather conditions, the variable degree day base is recommended to control for a greater portion of weather-related exogenous change along with a comparison group to address remaining weather-related change.

The alternative approach is variable degree-day, which entails the following steps:

1. Estimating each site-level regression and time period for a range of heating and cooling degree-day base combinations (including dropping heating and/or cooling components).
2. Choosing an optimal model (with the best fit, as measured by the coefficient of determination $R^2$ or CV(RMSE) within a specification and adjusted $R^2$, AIC, or BIC across models with different variables[13]) from among all of these models.

[13] Akaike information criteria and Bayesian information criteria are alternative measures for comparing the goodness of fit of different models.

The variable degree-day approach fits a model that reflects the specific energy consumption dynamics of each site. In the variable degree-day approach, the degree-day regression model for each site and time period is estimated separately for all unique combinations of heating and cooling degree-day bases, $\tau_H$ and $\tau_C$ across an appropriate range. This approach includes a specification in which one or both of the weather parameters are removed.

**Degree Days and Fuels**

For the modeling of natural gas consumption, it is unnecessary to include a cooling degree-day term. The gas consumption models tested should include the heating only (HO) and mean value options. Gas-heated households having electric water heat may produce models with negative baseload parameters. The models for these households should be re-run with the intercept (baseload) suppressed.

For the modeling of electricity, a model with heating and cooling terms should be tested, even if the premise is believed not to have electric heat or not to have air conditioning. Thus, for the electricity consumption model, the range of degree-day bases must be estimated for each of these options: a heating-cooling (HC) model, HO, cooling only (CO), and no degree-day terms (mean value).

**Degree Days and Set Points**

If degree days are allowed to vary:

- The estimated heating degree-day base $\tau_H$ will approximate the highest average daily outdoor temperature at which the heating system is needed for the day
- The estimated cooling degree-day base $\tau_C$ will approximate the lowest average daily outdoor temperature at which the house cooling system is needed for the day.

These base temperatures reflect both average thermostat set point and building dynamics, such as insulation and internal and solar heat gains.

The average thermostat set points may include variable behavior related to turning on the air conditioning or secondary heat sources. If heating or cooling are not present or are of a magnitude that is indistinguishable amidst the natural variation, then the model without a heating or cooling component may emerge the most appropriate model.

The site-level models should be estimated at a range of degree days that reflects the spectrum of feasible degree-day bases in the population. In general:

- A range of heating degree-day bases (from 55°F through 70°F) cover the feasible spectrum for single-family dwellings
- Cooling degree-day bases ranging from 65°F through 75°F should be sufficient.[14] (Note that the cooling degree-day base must always be higher than the heating degree-day base.)

[14] In both cases, it is important to remember that temperatures are based on average daily temperature and will be aggregated over a month or more of time.

A wider range of degree-day bases increases processing time, but this approach may provide better fits in some cases.

Plotting daily average consumption with respect to temperature provides insight into the inflection points at which heating and cooling consumption begin. However, mixed-heat sources may make a simple characterization of heat load such as this difficult.

For each premise, time period, and model specification (HC, HO, or CO), select as the final degree-day bases the values of $\tau_H$, and $\tau_C$ that give the highest $R^2$, along with the coefficients $\mu$, $\beta_H$, $\beta_C$ estimated at those bases. Models with negative parameter estimates should be removed from consideration, although they rarely survive the optimal model selection process.

#### *4.3.1.3 Optimal Models*

When the optimal model degree-day bases determined by the $R^2$ selection criterion are within the extremes of the temperature range tested, identify an optimal model. However, if the best-fitting model is at either extreme of the degree-day bases tested, this may not be the case. An extreme high- or low-degree-day base could indicate that the range of degree-day bases tested was too narrow, or it may reflect a spurious fit on sparse or anomalous data. If widening the degree-day base range or fixing anomalous data does not produce an optimal model within the test range, these sites should be flagged and plotted and the analyst should then decide whether the data should be kept in the analysis.

The practical response to degree-day base border solutions is to default to the fixed degree-day approach. In this case, the fixed degree-day bases could be fixed at the mean degree-day bases of all sites that were successfully estimated with a meaningful (non-extreme) degree-day base. Otherwise use 60°F for heating and 70°F for cooling. The NAC for these fixed degree-day base sites will still be valid, but the heating and cooling estimated parameters for these sites are potentially biased. This approach maximizes the information learned where the variable degree-day base approach works, but it defaults to the more basic approach where it fails.

Apply a consistent reliability criterion based on $R^2$ and the coefficient of variation (primarily for baseload-only models) to all site-level models. Ranking by $R^2$ is the simple way to identify the optimal degree-day choice within each specification (HC, HO, and/or CO). Use an appropriate statistical test to determine the optimal model among all of the different specifications (HC, HO, CO, and mean). The simplest acceptable selection rule is as follows[15]:

- If the heating and cooling coefficients in the HC model have p-values[16] less than 10%, retain both.
- Otherwise:
    - If either the heating coefficient in the HO model or the cooling coefficient in the CO model has a p-value of less than 10%, retain the term (heating or cooling) with the lower p-value.

[15] Adjusted R2, AIC or BIC are also used.
[16] A measure of statistical significance.

- If neither the heating nor the cooling coefficient has a p-value of less than 10% in the respective model, drop both terms and use mean consumption.
- For sites with no weather-correlated load or with a highly variable load, the mean usage-per-day may be the most appropriate basis for estimating normal annual consumption.

It is always possible to estimate a "best" model, but a number of caveats—such as those listed here—remain. Any interpretation of the separate heating and cooling terms from either the first stage of the stage-two model or the pooled model must recognize that these other uses are combined to some extent with heating and cooling.

- These models are very simple.
- Many energy uses have seasonal elements that can be confounded with the degree-day terms.
- During cold weather, the consumption of hot water, the use of clothes washers and dryers, and the use of lighting all tend to be greater.
- In summer, the refrigerator load and pool pumps tend to be greater.
- Internal loads from appliances, lighting, home office, and home entertainment reduce heating loads and increase cooling loads.
- Low-e windows and window films increase heating loads and reduce cooling loads.

To review, fixed degree-day base models can be used if the only information derived from the model is normalized annual consumption, because NAC is generally stable regardless of the degree-day base used. ***Fixed degree-day base models should not be used if the separate heating, cooling, or base components are to be interpreted and applied as such.***

#### *4.3.1.4 Step 2. Applying the Stage 1 Model*

To calculate NAC for the pre- and post-installation periods for each premise and timeframe, combine the estimated coefficients $\mu$, $\beta_H$, and $\beta_C$ with the annual normal- TMY[17] degree days $H_0$ and $C_0$ calculated at the site-specific degree-day base(s), $\tau_H$ and $\tau_C$. Thus, for each pre- and post-period at each individual site, use the coefficients for that site and period to calculate NAC. This example puts all premises and periods on an annual and normalized basis.

$$NAC = \mu * 365.25 + \beta_H H_0 + \beta_C C_0$$

The same approach can be used to put all premises on a monthly basis and/or on an actual weather basis. In instances where calendarization may be required, it may be preferable to use this approach to produce consumption on a monthly and actual weather basis, rather than using the simple pro-ration of billing intervals.

[17] Discussed in Section 4.4.6 in UMP *Chapter 17: Residential Behavior Evaluation Protocol*.

#### 4.3.1.5 Step 3. Calculating the Change in NAC

For each site, the difference between pre- and post-program NAC values (ΔNAC) represents the change in consumption under normal weather conditions.

### 4.3.2 Stage 2. Cross-Sectional Analysis

The first-stage analysis estimates the weather-normalized change in usage for each premise. The second stage combines these to estimate the aggregate program effect, using a cross-sectional analysis of the change in consumption relative to premise characteristics.

Three forms of the stage-two regression are recommended. Influence diagnostics should be produced for all stage-two regressions with outliers removed. Alternatively, some evaluators remove outliers based on data-dependent criteria such as 2.5 inter-quartile ranges from the median percent savings (established separately for the participant and comparison groups because they have different central tendencies and variances).

#### 4.3.2.1 Form A. Mean Difference of Differences Regression

As the most basic form of the stage-two regression, this approach produces the same point estimates as taking the difference of the average pre- and post-differences; however, it will produce slightly different standard errors as it assumes a common variance.

*Equation 2*

$$\Delta NAC_j = \beta + \gamma I_j + \varepsilon_j$$

where:

$\Delta NAC_j$ = change in NAC for customer j
$I_j$ = 0/1 dummy variable, equal to 1 if customer j is a (current-year) participant, 0 if customer j is in the comparison group
$\beta, \gamma$ = coefficients determined by the regression
$\varepsilon_j$ = regression residual.

From the fitted equation:

- The estimated coefficient $\gamma$ is the estimate of mean savings.
- The estimated coefficient $\beta$ is the estimate of mean change or trend unrelated to the program.

The coefficient $\beta$ corresponds to the average change among the comparison group, while the coefficient $\gamma$ is the difference between the comparison group change and the participant group change. That is, this regression is essentially a difference-of-differences formulation and can be accomplished outside of a regression framework as a difference of the two mean differences.

#### 4.3.2.2 Form B. Multiple Regression with Program Dummy Variables

This form allows for the estimation of savings for different measures or groups of measures. It may also include other available premise characteristics that can improve the extrapolation of billing analysis results to the full program population.

For whole-building programs, the typical savings magnitude can vary substantially across the different measures that may be implemented under the program. Regression with a single dummy variable produces a single average savings per premise across premises. With widely varying actions across premises, this average may not be well determined. Allowing for different average savings for different measure groups can result in a better estimated model. However, it's typically not possible to isolate the effects of each individual measure. It's most effective then to include only a handful of measure groups, such as one to three large-impact measures individually, plus all others as a group.

*Equation 3*

$$\Delta NAC_j = \Sigma_q \beta_q x_{qj} + \Sigma_k \gamma_k I_{kj} + \varepsilon_j$$

where:

$I_{kj}$ = 0/1 dummy variable, equal to 1 if customer j received measure group k in the current year, 0 if customer j is in the comparison group and/or did not receive measure group k.

$x_{qj}$ = value of the characteristics (square footage, number of occupants, etc.) variable q for customer j. Let $x_{0j}$, the first term of this vector, equal 1 for all premises, so that $\beta_0$ serves as an intercept term.

$\beta_q$, $\gamma_k$ = coefficients determined by the regression.

From the estimated equation:

- The estimated coefficient $\gamma_k$ is the estimate of mean savings per participant who received measure group k.
- The coefficient $\beta_q$ is the estimate of mean change or trend unrelated to the program per-unit value of variable $x_q$.

This form may be used with any of the following:

- Multiple characteristics variables $x_q$ and a single measure dummy variable I
- Multiple dummy variables $I_k$ and a single characteristics variable x (other than the intercept)
- Only an intercept term (no premise characteristics) and a single dummy variable, I.

If only an intercept term and a single dummy variable are used, this form reduces to the first model type. For this type of regression to be meaningful, it is essential that the characteristics variables ($x_q$) are obtained in a consistent manner for both the participants and the comparison group. For many programs, if the comparison group is future or prior participants, these variables may be obtained from tracking data collected the same way across the program years.

#### 4.3.2.3 Form C. Statistically Adjusted Engineering Regression with Program Dummy Variables

This form adds the expected savings into the regression specification. If the expected savings from the tracking data are more informative than the simple indicator variable used in the previous specifications, then this approach should have greater precision. The model structure assumes an additive relationship between multiple measures which may not reflect interactive effects. Measure combinations can be parameterized to capture interaction effects explicitly.

*Equation 4*

$$\Delta NAC_j = \Sigma_q \beta_q x_{qj} + \Sigma_k \gamma_k I_{kj} + \Sigma_k \rho_k T_{kj} + \varepsilon_j$$

where:

$T_{kj}$ = tracking estimate of savings for measure group k for current-year participating customer j, 0 for customer j in the comparison group

$\beta_q, \gamma_k, \rho_k$ = coefficients determined by the regression

From the fitted equation:

- The mean program savings must be calculated using the coefficients on both the participation dummy variables and the tracking estimates of savings. That is, the estimated mean program savings for measure group k with mean tracking estimate $T_k$ is:

  $$S_k = \gamma_k + \rho_k T_{k\_}$$

- The coefficient $\beta_q$ is the estimate of mean change or trend unrelated to the program per-unit value of variable $x_q$.

This form may be used with any of the following:

- Multiple characteristics variables $x_q$ and a single measure group
- Multiple measure groups k and a single characteristics variable x (other than the intercept)
- Only an intercept term, no premise characteristics and a single measure group.

For each measure group k in the model, both the dummy variable $I_k$ and the tracking estimate $T_k$ should be included.

A simpler Statistically Adjusted Engineering (SAE) form that omits the participation dummy variable has the nominal appeal of the coefficient $\rho_k$ being interpreted as the "realization rate," the ratio of realized to tracking savings. However, inclusion of the tracking estimate without the corresponding dummy variable can lead to understated estimates of savings due to errors from omitted variables bias.

If the tracking estimate of savings is a constant value for all premises, the inclusion of the tracking estimate will not improve the fit. Moreover, if the tracking estimates vary but in ways that are not well correlated with actual savings, the fit will tend to be poor, with some savings

coefficients not significant and others not realistic. As for the multiple-dummy variable approach Form B described in Section 4.3.2.2, the SAE approach works best if the number of separate measure groups k is kept small. If the SAE approach does not produce meaningful results, the multiple- or single-dummy-variable version is preferred.

### 4.3.3 Instrumental Variables Regression

Instrumental variables (IV) regression addresses a potential bias in the basic regression that can arise if the tendency to participate is correlated with the change in consumption unrelated to the program. Such a correlation will tend to exist if any of the following are true:

- Free-ridership is present, at a non-negligible level
- The comparison group includes customers for whom the program measures aren't applicable
- Customers tend to participate in the program at times when they're taking other actions or have life events that generally tend to increase (or that generally tend to decrease) consumption.

Measure applicability is a particularly a concern when free-ridership is present. Customers for whom the program measures wouldn't apply or wouldn't make sense have zero natural adoption and don't participate in the program. Thus, the inclusion in the comparison group of customers who couldn't benefit from the program measures exacerbates the mismatch between the participant and comparison groups' rates of natural adoption.

The IV method adds an additional step to the regression process. Specifically, a model that predicts participation as a function of observable variables is fit. If an RED is used, the encouragement dummy variable becomes a predictor in the participation model. Common forms of the participation model include a logit or probit.

The fitted model is then used to calculate the participation probability for each customer in the analysis, and this participation probability is substituted for the participation dummy in Eq. 2 or 3. In the simplest form with an RED, the encouragement variable is the only predictor in the participation equation, and Eq. 2 with the substitution of predicted for observed participation is used for the analysis. In this form, the result is equivalent to the difference of differences LATE estimator described in Section 3.5.

Conditions for the participation model specification include the following:

1. It should include all the explanatory variables $x_q$ included in Eq. 3 above.
2. It should include one or more variables that DON'T directly affect energy consumption but DO affect participation.
3. If there are any additional (observable or unobservable) consumption drivers that are left out of the consumption equation, the participation model predictors must be unrelated to any of these omitted terms.

The IV approach may be used with or without an RED. However, without an RED, it is difficult for the 2nd participation model condition to be satisfied. It also may be difficult to get good predictive power for the model. If the participation model has weak ability to separate high- from

low-participation customers, the IV analysis will tend to yield savings estimates with high variance.

The basic IV analysis cannot provide an unbiased net savings estimate for the main program when free-ridership is present. However, the IV analysis with RED does control for unobservable factors that affect naturally occurring change but don't also affect the net savings a customer will have if they join the program. In many whole-building programs there is a tendency to join a program at a time of major renovation, which tends to increase consumption. On the other hand, customers might choose to join a pay-for-performance program if they anticipate household changes that will tend to bring consumption down. The RED can eliminate the bias due to factors such as these that tend to work in a particular direction for a particular program. If the three conditions noted above for the participation model are met, the average effect of such factors, for a given participation probability, is zero. As a result, there is no confounding of these unrelated changes with the estimated participation effect.

Moreover, as noted in Section 3.5, when free-ridership is present, the LATE estimate from the RED with basic IV analysis does give the net savings per incremental participant attributable to the incremental encouragement. Since free-ridership is likely to be lower among those who require supplemental encouragement to join, this LATE estimate can be regarded as an upper bound on the base program net savings.

The use of RED and IV methods is discussed more fully in Goldberg et al. (2017). That work also describes an additional method that can potentially provide an unbiased estimate of net savings for the main program, using an extension of the basic IV method. While that method is promising, further empirical work is needed before specific recommendations can be offered for its use.

### 4.3.4 Choosing the Stage-Two Regression Form

The mean difference-of-differences regression estimate (described earlier) is recommended if the following three conditions are met:

- Only overall average program savings is to be estimated, rather than separate savings for different groups of measures
- Factors that may be associated with differences in the magnitude of the non-program trend (such as square footage) are the same on average for the current-year participant group as for the comparison group
- More precise estimates are not required, or additional data that could yield a more accurate estimate are not available.

The second general model, Form B (Multiple Regression with Program Dummy Variables), is recommended if:

- Either (a) separate savings estimates are desired for different groups of measures, ***or*** (b) factors that may be associated with differences in the magnitude of the non-program trend (such as square footage) are not the same on average for the current-year participant group as for the comparison group

- Informative tracking estimates of savings are not available.

The third general model, Form C (SAE Regression With Program Dummy Variables), which incorporates a tracking estimate of savings, is preferred when there are both an informative tracking estimate of savings ***and*** an interest in more refined estimates than can be obtained with the simplest model version.

Forms B and C make it possible to extrapolate the consumption data analysis results back to the full tracking data based on measure-level results. This may be of particular importance, depending on the extent and nature of the attrition of tracking data sites out of the analysis dataset.

If an informative tracking estimate is not available but there are characteristics variables likely to correlate with savings, then a proxy for savings constructed from these characteristics variables can be substituted for the tracking estimate. Proxies that may usefully inform a second-stage model include count of light bulbs and the square footage of installed insulation.

## 4.4 Pooled Fixed-Effects Approach

The pooled approach can be specified either with a comparison group or with multiple years of participants. With a comparison group, the pooled model is a pooled version of the 2-stage approach discussed above. With multiple years of participants included in the pooled model the later participants are implicitly performing the function of comparison group. The comparison group approach offers a more straightforward specification and is the focus of this section.[18]

The basic structures of the site-level and the second-stage consumption data model are effectively combined in the pooled approach. All monthly participant and comparison group consumption data (both pre- and post-installation) are included in a single model. This model has:

- A site-level fixed-effect component (analogous to the site-level baseload component)
- A monthly fixed effect
- A participant group indicator variable (absorbed into the site-level fixed effect when not interacted with other variables)
- A post-installation indicator variable capturing the change in the post-installation period across participant and comparison groups
- A participant-post combined indicator that captures the savings estimate
- Heating and cooling components interacted with the participant indicator variable, the post-installation indicator variable, and the participant-post combined indicator variable.

### *4.4.1 Recommended Form of Pooled Regression*

An example pooled model equation is as follows:

[18] The discussion in the section parallels discussion in Section 4.3.6 of *Chapter 17: Residential Behavior Protocol*. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures.

*Equation 5*

$$E_{jm} = \mu_j + \phi_m + \beta_H T_j H_{jm} + \lambda P_{jm} + \lambda_H P_{jm} H_{jm} + \gamma T_j P_{jm} + \gamma_H T_j P_{jm} H_{jm} + \varepsilon_{im}$$

where all variables have already been defined except for these:

| | | |
|---|---|---|
| $\mu_j$ | = | Unique intercept for each participant j |
| $\phi_m$ | = | 0/1 Indicator for each time interval *m*, time series component that track systematic change over time |
| $P_{jm}$ | = | 0/1 Indicator variable for the post-installation period for both treatment and comparison groups. |
| $\beta_H,\lambda,\lambda_H,\gamma,\gamma_H$ | = | coefficients determined by the regression |

This specification only includes heating terms ($H_{jm}$), as if for a gas analysis; however, analogous cooling terms should be included for an electric pooled model.

The parameter interactions that include only the variable $P_{jm}$ capture the post-period effect for both participants and the comparison group. The parameter interactions that only include $I_j$ control for differences between the participant group and comparison group in the pre-period.[19] The parameter interactions with both $P_{jm}$ and $I_j$ represent the post period effect on participants given the other interactions. This specification is the regression version of the difference in difference approach.

The mean program savings is calculated using the following equation:

$$S = \gamma * 365.25 + \gamma_H H_0$$

where:

$H_0$ = TMY degree days at the base for the regression

The pooled regression can also be specified as an SAE model.

### 4.4.2 Choice of Pooled Form

The pooled approach features a simplified weather-normalization structure compared to the site-level modeling in the two-stage approach. All buildings are characterized by a mean heating and/or cooling slope calculated from a fixed degree day base. In addition, the panel structure requires regression errors to be clustered at the building level to address the lack of independence of consumption across month within a building. The primary advantage of the pooled structure is the avoidance of site-level modeling altogether. In general, the pooled approach will also provide estimates with a higher precision, even after clustering, due to the increased size of the dataset.

## 4.5 Data

### 4.5.1 Basic Data Preparation

Before a consumption data analysis can be performed, the following activities must be done. The details of these steps are provided later in this section.

[19] The mean difference between the two groups is accounted for in the site-level fixed effect, $\mu_j$.

1. ***Obtain program tracking data for current year participants.*** The tracking data should identify what program measures were installed and on what date. These data may also include some customer or building characteristics.
2. ***Identify the comparison group customers.*** Obtain tracking data for these customers if they are previous or future participants, so as to assure that all comparison group consumption data are either fully pre- or fully post-participation in the program.
3. ***Obtain consumption data files from billing records for each building in the analysis.*** This may require mapping participant account numbers to premise accounts. Buildings with occupant turnover during the evaluation period should be assessed separately and may warrant removal from the analysis.
4. ***Screen and clean the consumption data*** as described in *Data Requirements and Collection Methods,* Section 4.5.2.
5. ***Convert the billing records for each meter reading interval*** to average consumption-per-day for each premise.
6. ***Identify the pre- and post-periods for each premise in the analysis.*** Based on the installation dates, the pre- and post-installation periods are defined for each participant to span approximately 12 months before and approximately 12 months after installation. The billing interval or intervals during which the measure was installed for a particular participant include both pre- and post-installation consumption days. These transitional billing intervals should be excluded from the analysis. (The excluded billing intervals are referred to as the blackout intervals for that participant.) The post period is identified with 0/1 dummy variable.
7. ***Identify the nearest weather station associated with each premise in the analysis.*** The utility may maintain a weather station look-up for this purpose, so use that if it is available. In general, weather station assignments should consider local geography rather than simply selecting the nearest station. For example, in California, the weather station should be in the same climate zone as the home. Also, consider all significant elevation differences in the station assignment**.**
8. ***Obtain daily temperature data from each weather station*** for a period that matches the consumption data.
9. ***Determine for each weather station the actual and normal heating and cooling degree days*** for degree day base temperatures—from 55°F through 75°F—for each day included in the analysis, as is detailed in the *Data Requirements and Collection Methods* section below.
10. ***Calculate average daily degree days*** for the exact dates of each bill interval in the consumption data.

### 4.5.2 Data Requirements and Collection Methods

A consumption data analysis requires data from multiple sources:

- Consumption data, generally from a utility billing system
- Program tracking data
- Weather data.

This section describes the required data for a whole-building retrofit billing analysis and the steps for using these data correctly.

#### 4.5.2.1 Consumption Data

The consumption data used in a consumption data analysis are generally stored as part of the utility billing system. Because these systems are used by evaluators relatively infrequently, recovering consumption data from the system can be challenging. To obtain the needed data, prepare a written request specifying the data items, such as:

- Unique site ID
- Unique customer ID
- Read date
- Consumption amount
- Read type (indicating estimated and other non-actual reads)
- Variables required to merge consumption data with program tracking data
- Location information or other link to weather stations
- Customer tenancy at the premise (the tenancy starting and ending dates)
- Other premise characteristics available in the utility customer information system, including dwelling type, heating or water heating fuel indicators, or participation in income-qualified programs.

It is essential to establish the unique site identifier with the help of the owner of the data at the utility. Note that the unique site ID specifies the unit of analysis. Usually, a combination of customer and site/premise ID identifies a particular location with the consumption data for the occupant.

The primary data used for a consumption data analysis are the consumption meter reads from the utility revenue meter, and these readings are typically taken monthly or bimonthly for gas and electric utilities in the United States. The consumption data are identified with specific time intervals by a meter read date and either a previous read date or a read interval duration. Average daily consumption for the known monthly or bi-monthly time interval is calculated by combining these data, which then serve as the dependent variable for all of the forms of consumption data regression.

The remaining requested variables serve one of three purposes:

- Linking the consumption data with other essential data sources (such as program tracking data and weather data)
- Providing information that facilitates the cleaning of the consumption data
- Providing data for characterizing the household so as to improve the quality of the regression models.

**Consumption Data Preparation**

Consumption data received from the service provider are likely to be subject to some combination of the following issues, which are provided here as a checklist to be addressed. It is almost impossible to prescribe definitive rules for addressing some of these issues, as they arise

from the unique conditions of each billing system. This list represents the common issues encountered in consumption data and provides basic standards that should be met. The general goal should be to limit the analysis to intervals with accurate consumption data with accurate beginning and ending dates.

- ***Zero reads.*** Zero electric reads are rare and usually indicate outages, vacancy, or other system issues. Zero gas reads, however, are more common. Infrequent zeros in an electric data series can be ignored, as can zero reads in gas series during the non-heating months. Sites with extensive electric zero reads or zero gas reads during the heating season should be identified and removed.
- ***Extreme data.*** Sites with extreme reads should be removed unless evidence indicates that high-level usage patterns are typical. Atypical extreme spikes are frequently the result of meter issues, so it is best to omit them from the analysis. For smaller populations: (1) Plot and review consumption levels above the 99$^{th}$ percentile of all consumption levels. Alternatively, flag points that are more than three inter-quartile ranges away from the median consumption. (2) Develop realistic consumption minima and maxima for single-family homes. The decision rule should be applied consistently to the participant and comparison groups.
- ***Missing data.*** Missing data should be clearly understood. Some instances are self-explanatory (pre- or post-occupancy), but many are not, and these require an explanation from the utility data owner. Because true missed reads are generally filled with estimations, missing data in the final consumption indicate an issue worth exploring.
- ***Estimated reads.*** A read type field, available from most billing systems, indicates whether a consumption amount is from an actual read or some form of system estimate. Any read that is not an actual read should be aggregated with subsequent reads until the final read is an actual read. The resulting read will cover multiple read intervals, but the total consumption will be accurate for the aggregated intervals.
- ***First reads.*** The first read available in a consumption data series may correct for many previous estimated reads. Each site data series used for the analysis should begin with a consumption value that is a confirmed single-read interval. This entails removing all leading estimated reads from the series and then removing one additional, non-estimated leading read from each site data series.
- ***Off-cycle reads.*** Monthly meter reading periods that span fewer than 25 days are typically off-cycle readings, which typically occur due to meter reading problems or changes in occupancy. These periods should be excluded from the analysis.
- ***Adjustments.*** Adjustment reads may either be single reads that are out of the normal schedule or reads combined with a normally scheduled read. Adjustments may be indicated by the read-type variable, or they may appear, for instance, as a consistent spike in December reads. Adjustments correct a range of errors in previous consumption data in a one-time, non-informative way. Unless the magnitude of the adjustment is small, such adjustments necessitate the removal of prior data from a site and may require the complete removal of the site if enough data are compromised.

- ***Overlapping read intervals.*** Because overlapping read intervals may indicate an adjustment or a data problem, they should be discussed with the data owner. If these read intervals undermine the consumption-weather relationship, then the site must be removed.
- ***Multiple meters.*** Although having multiple meters is rare in single-family housing, this situation does exist. When multiple meters are read on the same schedule, as is usually true for such residences, the meter reads for the same home should be aggregated to the household level for each meter reading interval.

As consumption data analysis is generally applied to the full population of a program, dropping small percentages of sites is unlikely to affect the results. However, if the number of removed sites increases beyond 5%, it is worth considering whether the issues causing removal are possibly correlated with some aspect of program participation and/or savings. This issue could lead to biased results. If removal is greater than 5%, then the analysis should include a table that compares the analysis group to the program participant population on available data (such as house characteristics, program measures, and pre-retrofit usage).

#### 4.5.2.2 Weather Data

Weather data are used in the consumption data analysis in two ways:

- In models that relate consumption to weather, the observed weather data are matched to the meter read intervals to provide predictor variables.
- The model estimated with actual weather is calculated at normal-year weather levels to provide usage and savings in a normal or typical year.[20]

Use either primary National Oceanic and Atmospheric Administration (NOAA) or weather stations managed by the utility (and trusted by utility analysts) as the source for weather data. Some utilities maintain weather series (both actual and normal/TMY) for internal use, and it is generally best to use a utility's weather resources to produce evaluation results that are consistent with other studies within the utility. Many utilities are choosing to use norms constructed from fewer than 30 years, as are the standard NOAA norms.

A consumption data analysis requires both actual and normal (or TMY) weather data from a location near each premise. The actual weather data must match the time interval of each meter reading interval. Both actual and normal/TMY weather data used for each site should come from the same weather station. Only annual TMY degree days are required for annual analysis results. This protocol recommends calculating the annual monthly normal degree days for the purpose of plotting model fit values.

#### 4.5.2.3 Weather Data Preparation

Depending on the source, weather data may need additional preparation. Limited missing data can be filled by the simple interpolation. If the amount of missing data is sufficient to trigger

[20]NOAA produces 30-year normal weather series composed of average temperature for each hour over the time period. These norms are updated every decade. NREL produces TMY data series. These data are not average values but a combination of typical months from years during the time period. The TMY data also cover a shorter time period.

concern regarding a weather data source, consider using a more distant but more complete weather station as an alternative.

Create a graph to identify anomalies, gaps, and likely data errors. Weather data issues tend to be obvious visually. Missing data and technical failures look very different than naturally random weather patterns. For each weather station used in the analysis, plot the following information over the analysis time span: minimum, maximum, and average temperature versus day of year. If multiple weather stations are used across a large region, plot the different stations on a single graph.

#### 4.5.2.4 Tracking Data

The program tracking data provide the participant population, the installation date or a proxy such as paid date, and the number and type of measures for which savings are claimed. Frequently, the original consumption data request is made based on the population defined by the tracking data. Additional information in the tracking database may serve as a resource for other elements of the analysis:

- If a variety of measures were installed and there is a sufficient mix of different combinations of measures, it may be possible to develop savings estimates for some individual measures. In this situation, focus the evaluation on the measures with greater expected savings for separate estimates of savings.
- The date of a measure's installation both provides the date at which the change in consumption took place ***and*** identifies the billing interval(s) that will be blacked out. The tracking database, however, may contain the installation confirmation date, the date of payment, or some other date loosely associated with the time at which consumption actually changed (rather than the explicit installation date). The evaluator should consult with the program staff to determine what the different recorded dates refer to and when actual installation could have occurred in relation to these dates.

  Also, it may be necessary to black out multiple billing periods. Multiple installation dates at the same site may require a longer blackout period or may make the site untenable for simple pre-post analysis. If the blackout period does not encompass the dates of all program-related changes to consumption, then the pre-post difference will be downwardly biased.
- The tracking data may also be a useful resource regarding the characteristics of participant homes. Frequently, program databases capture home square footage, number of floors, existing measure capacity, and efficiency. These data are primarily useful in the pooled approach if they are only available for current participants.
- Tracking data from previous years may be used to define a control group for a two-stage analysis.

### 4.5.3 Analysis Dataset

Using the account numbers in the two datasets, the final analysis dataset combines the tracking data and the consumption data with the weather data. Weather data are attached to each consumption interval, based on the days in a read interval. The combined data have a sum of the

daily degree-days for each unique read interval, based on start date and duration. If the variable degree-day base approach is used, this process must be repeated over the range of heating and cooling degree-day bases. To produce average daily consumption and degree days for that read interval, the read interval consumption and degree-day values are divided by the number of days in the interval.

Because of the complication of matching weather to all the unique read intervals, some evaluators resort to calendarized data.[21] Except in special cases, calendarization should not be used for this kind of analysis because it undermines the direct matching between consumption and degree days that is the basis of consumption data analysis. Multiple meter and multifamily analyses are examples of situations where calendarization may be the only way to aggregate data series on different schedules.

#### 4.5.3.1 Analysis Data Preparation

A number of additional data preparation steps are required when the three data sources (tracking, billing, and weather) have been combined. These limit the analysis data to only the data to be included in the model.

- ***Participant Data Only.*** Confirm that the consumption data in the analysis dataset is only for the household occupant who participated (or will participate) in the program.
- ***Blackout Interval.*** Remove from the regression the full read interval within which the installation occurred. If the installation timing is not explicitly indicated in the tracking system—or if installation occurred in stages over several weeks or had ramp-up or ramp-down effects—it may be necessary to extend the blackout interval beyond a single read interval.
    - For a single, relatively simple measure (such as a furnace), a single blackout month is sufficient.
    - For more complex installations (longer-term single installations or multiple installations), a multiple-month blackout may be more appropriate.

  The change in consumption will be biased in a downward direction if part of the transition interval is included as either pre- or post-installation typical consumption. In most instances, the only negative aspect of increasing the blackout interval is the corresponding decrease in either pre- or post-installation readings.
- ***Sufficient Data for a Site.*** Count the number of data points in the pre- and post-blackout periods for each individual site consumption data series. To create a view of the classic seasonal consumption data patterns, plot a representative sample of daily average consumption data by read date. Daily average consumption plotted by temperature replicates the underlying structure of the consumption data analysis. Plotting the estimated and actual monthly values in both formats is the most effective way to identify unexpected issues in the data and to reveal issues related to model fit.

  Ideally, a full year of consumption data is available for each site for the pre- and post-blackout periods.

[21] Calendar month consumption is estimated as a weighted average of the bill readings that cover that month.

- For individual site analysis of electric consumption, a minimum of nine observations spanning summer (July and August), winter (January and February), and shoulder seasons are recommended for each site in each time period (pre- and post-installation). For gas consumption, six observations spanning at least half of a winter and some summer are the minimum.
- For a pooled analysis, sites with fewer observations or fewer seasons represented can be included (a minimum of six in each period). However, it is important to have all seasons represented in both time periods and across all premises in the pooled model.
- Bimonthly data provide a particular challenge for consumption data analysis. In a year of data, all seasons are represented, but the number of data points is halved. For analysis of gas consumptions, a minimum of one year each of pre- and post-installation data is essential. For analysis of electric consumption, two years each of pre- and post-blackout data are better.

# 5 Looking Forward

As discussed in Section 2.1, more granular AMI data are increasingly available to evaluators pursuing consumption data analysis. These data bring new opportunities and new challenges to evaluation. While the more granular data offer the possibility of estimating the peak period kW effects and time-differentiated energy efficiency impacts (kWh) for a program, they also increase the breadth and complexity of modeling approaches and the computing power required to produce results. Also, although the granularity of available consumption data is increasing, the other data available for inclusion in a typical evaluation model—tracking data, weather, etc.—remain mostly the same.

Whole building evaluation will benefit substantially by incorporating the learning from site-level modeling efforts that have been pursued for years in the commercial sector where interval data have been available, as well as from demand response/direct load control modeling efforts that have used both end-use and whole-building data for the purpose of modeling short term load curtailments. A protocol addressing the use of AMI data for consumption data analysis will contend with almost all the issues put forward in this chapter as well as the additional challenges revealed with the more granular data—the diurnal patterns combined with the unique thermal dynamics of each building.

As consensus is reached on the best practices in the use of AMI data for consumption data analysis, an additional chapter, or a substantially expanded version of this chapter, will be needed to capture these practices.

# 6 References

Fels, M.F., ed. (February/May 1986). Energy and Buildings: Special Issue Devoted to Measuring Energy Savings: The Scorekeeping Approach. (9:1&2).

Fels, M.F.; Kissock, K; Marean, M.A.; and Reynolds, C. (January 1995). PRISM (Advance Version 1.0) Users' Guide. Center for Energy and Environment Studies. Princeton, New Jersey.

Goldberg, M. L., Fowlie, M.; Train, L.; and Agnew, G.K. (2017, in revision). A White Paper: Mitigating Self-Selection Bias in Billing Analysis for Impact Evaluation. DNV GL, submitted to PG&E.

Goldberg, M. L., Fowlie, M.; Train, K.; and Agnew, G.K. (2017). Not Just Another Pretty Formula: Practical Methods for Mitigating Self-Selection Bias in Billing Analysis Regressions. Proceedings of the International Energy Program Evaluation Conference. Baltimore, Maryland.

Jacobson, D. (2017). *Chapter 5: Residential Furnaces and Boilers Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68561. http://www.nrel.gov/docs/fy17osti/68561.pdf

Randazzo, K.; Ridge, R.; and Wayland, S.(2017, in revision). Observations on Chapter 8 of the Uniform Methods Project: A Discussion of Comparison Groups for Net and Gross Impacts. Opinion Dynamics, submitted to PG&E.

Stewart, J. and Todd, A. (2017). *Chapter 17: Residential Behavior Protocol. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68573. http://www.nrel.gov/docs/fy17osti/68573.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

# 7 Resources[22]

ASHRAE. (TBD). *Guideline 14-2002R*. (Revision of *Guideline 14*, currently in process).

ASHRAE. (2010). *Performance Measurement Protocols for Commercial Buildings*.

ASHRAE. (2002). *Guideline 14-2002 Measurement of Energy and Demand Savings*.

ASHRAE. (2004). *Development of a Toolkit for Calculating Linear, Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models*. ASHRAE Research Project 1050.

[22] Some resources recommended by ASHRAE.