Whole-Building Retrofit Evaluation Protocol

1 Measure Description 1
2 Application Conditions of Protocol 1
3 Savings Calculations 3
3.1 General Approaches 3
3.1.1 Two-Stage 3
3.1.2 Pooled 3
4 Comparison Group Specification 5
4.1.1 Self-Selection and Freeridership 6
4.2 Recommendations by Program Characteristics 8
4.3 The Full-Year Specification 9
4.4 The Rolling Specification 9
4.5 The Two-Stage Approach 12
4.5.1 Stage 1. Individual Premise Analysis 12
4.5.2 Step 2.Applying the Stage 1 Model 17
4.5.3 Step 3.Calculating the Change in NAC 17
4.6 Stage 2. Cross-Sectional Analysis 17
4.6.1 Recommended Forms of Stage-Two Regression 17
4.6.2 Choosing the Stage-Two Regression Form 20
5 Pooled Fixed-Effects Approach 21
5.1 Recommended Form of Pooled Regression 21
5.1.1 Choice of Pooled Form 22
6 Measurement and Verification Plan 24
6.1 IPMVP Option C 24
6.2 Verification Process 24
6.3 Data Requirements and Collection Methods 24
6.3.1 Billing Data 24
6.3.2 Weather Data 27
6.3.3 Tracking Data 27
6.4 Analysis Dataset 28
6.4.1 Analysis Data Preparation 29
7 Sample Design 31
7.1 Program Evaluation Elements: Considerations for Other Program Types and Conditions 31
7.2 Alternative Comparison Group Specifications 31
8 References and Resources 33

Whole-Building Retrofit Evaluation Protocol

Ken Agnew, Mimi Goldberg, DNV KEMA

Whole-building retrofit programs focus on a building's energy performance overall. Typically, these programs involve installing a mix of energy-efficiency measures that, in combination, reduce the total energy consumption of a house or facility. Examples of whole-building retrofit programs are home weatherization, Home Performance with ENERGY STAR®, and many low-income programs. While whole-building retrofit programs generally target residential buildings, they may also target small commercial buildings.

1 Measure Description

Because whole-building retrofits involve the installation of multiple measures, the estimation of the total savings requires a comprehensive method for capturing the combined effect of the installed measures. The general method recommended for this type of program is a billing analysis—the analysis of consumption data from utility billing records. This method is consistent with the recommended International Performance Measurement and Verification Protocol[1] (IPMVP) Option C, Whole Facility. Option C is designed in part to address evaluation conditions that occur with a whole-house retrofit program.

The billing 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 billing analysis can be an effective evaluation technique for whole-house retrofit programs, and it addresses both how and when billing analysis should be used.

2 Application Conditions of Protocol

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 evaluation methods noted in this chapter are applicable when all of the following are true:

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

[1] International Performance Measurement and Verification Protocol (IPMVP), which is considered the gold standard for evaluating energy-efficiency programs.

- The baseline for determining savings is the condition of the participating building before the retrofits were made, rather than the standard efficiency of the new equipment.
- There is sufficient consumption data available—in the form of monthly or bi-monthly utility billing records— for the participants[2].
- (Optional) Consumption data are available for the same timeframe as for the participants for one or more of the following groups: (1) previous participants—those who took part in the program before the timeframe of the current evaluation; (2) subsequent participants; or (3) those who are on a list for future participation in the program.

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 the Furnace Boiler protocol uses a billing analysis result and addresses the baseline issue described in the third bullet above.[3]

---

[2] Daily consumption data are now available from some billing systems. From the perspective of billing analysis evaluation, such data are a finer-grained form of the same basic data. The methods discussed here are primarily applicable to daily consumption data. There are issues unique to daily data, and one obvious concern is increased serial correlation in the modeling process and the resulting artificially low standard errors. (Note that this protocol also does not explore the additional opportunities that are available with the finer grain data.)

[3] As discussed under Section 7 of the Introduction chapter to the UMP Report, small utilities (as defined under the SBA regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

3 Savings Calculations

Because whole-house retrofit programs install multiple measures, the estimation of the total savings requires a comprehensive method for capturing the combined effect of all of the installed measures. The general approach recommended for this type of program is a billing analysis.

3.1 General Approaches

Two general billing analysis approaches are described here: "two-stage" and "pooled."

3.1.1 Two-Stage Approach

This approach is recommended in cases where there are: (1) a valid comparison group, and (2) sufficient consumption data for each building in the analysis. The Two-Stage method[4] consists of these activities:

- In ***Stage 1, the weather-normalized annual consumption (NAC) is estimated separately for each building*** in the analysis for both the pre- and post-program periods. The weather normalization for each building and period relies on a longitudinal regression analysis. Observations in these regressions correspond to usage over different bill intervals (typically, months) for the same building. For participants, the difference between the building's pre- and post-program NAC represents the program-related change in consumption plus exogenous change. For non-participants the pre-post difference represents only exogenous change.
- In ***Stage 2, a cross-sectional analysis is conducted on the Stage 1 output*** to isolate the aggregate program-related change from the observed changes in consumption. Depending on how the regression equation is specified, observations in the second-stage analysis are either the change in NAC for different customers, or the separate pre- and post-program year NACs for different customers and pre- and post- periods.

3.1.2 Pooled

The pooled approach combines all participants and time intervals into a single regression analysis. This is also referred to as a "time-series cross-sectional analysis" because its observations vary both across time and across individual buildings.

The pooled approach is appropriate under most scenarios described here, but it is particularly recommended when either of the following is true:

- There is not a valid, separate comparison group;
- The goal is to measure an average effect over multiple program years.

The conditions for obtaining reliable results in these situations are described in a later section under the heading "Pooled Fixed Fixed-Effects Approach."

For the evaluation of a whole-house retrofit program, the following are recommended:

---

[4] The two-stage billing analysis is not the same as the econometric "2-Stage Least Squares" regression method.

1. Use past and future (or “pipeline”) participants as the comparison group for the current program year. (See the details in the next section.)

2. Use a Two-Stage Approach unless the consumption data are too limited to produce good normalization models for individual buildings (as discussed below). In that case, use the pooled method.

3. Interpret savings carefully so they can be adjusted for freeridership as necessary. (Most billing analysis results are either gross savings or fall somewhere in between net and gross.) The following section discusses this issue.

The comparison group specification is described next, followed by the Two-Stage Approach using this comparison group. Then the pooled analysis using the same data is described.

4 Comparison Group Specification

Choosing the right comparison group is of central importance for a successful billing analysis. The goal of a billing analysis is to measure the change in building energy consumption from the pre-program period to the post-program period without including the effect of natural changes in consumption not due to the program. The comparison group makes it possible to remove these other changes in consumption—referred to here as exogenous changes—resulting from changes in fuel prices, general economic conditions, natural disasters, etc.[5]

The optimal evaluation scenario for a billing analysis is a randomized controlled trial (RCT) experimental design. This is essentially 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.

The control group sets the standard by which billing analysis comparison groups should be assessed. For an RCT, a sampling of eligible participants is randomly assigned to one of two groups before the program installations **(**treatment). This assures that the two groups—treatment and control—are probabilistically similar in every respect except for the offer of program treatment. The basic structure of this process is a "difference of differences." 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 represents the program-related change plus exogenous change.
- For the control group, the pre-post difference represents only exogenous change.

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. At present, in the context of energy-efficiency programs, true RCT is rare outside of certain types of behavioral programs.[6] The approach remains the gold standard, however, and provides a good illustration of the ideal characteristics of a control group.

Where a program is not designed as an RCT, 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 perform function as an experimental control group.

---

[5] While weather-related change is a form of exogenous change, it is controlled for in the models.

[6] There are multiple reasons why RCT has not been more widely employed. Until recently, evaluation concerns have been less likely to drive program planning. Also, RTC requires denying or delaying participation to a subset of the eligible, willing population and, under some approaches, it involves giving services to people who either do not want them or may not use them. The importance of RCT to the evaluation process is motivating program administrators to consider incorporating RCT into their program structures more frequently.

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. 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 non-participants.

In theory, many important characteristics can be controlled for; 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 freeridership. One of the many possible characteristics that could define a program participant is the intent to perform energy-efficiency activity regardless of program support. 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.

4.1.1 Self-Selection and Freeridership

The interaction between self-selection and freeridership 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 is freeriders.

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

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 energy-efficiency 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 energy efficient-inclined population. Customers and contractors inclined toward energy efficiency have little reason not to take advantage of the rebates. This is likely to lead to an over-representation of natural adopters in the participant population, as compared to the general incidence in the population. This, then, affects in multiple ways the level of savings and freeridership that will be measured by the billing analysis.

- First, any comparison group developed after the fact from those who chose not to participate will tend to have a lower percentage of energy-efficient furnace installers (in this example) than would a true control group. To the extent that this is the case, the comparison group will not control for the full extent of natural energy-efficient furnace installations had the program not been in place.
- Second, the treatment group includes a higher proportion of natural energy efficient adopters than the general population, due to self-selection into the program. These households increase the freeridership rate beyond the natural level of natural adopters in the eligible population.
- Finally, the more general concerns regarding self-selection are still present. Because of their natural inclination to adopt energy efficient, the participants are likely to exhibit different energy-consumption characteristics than the general population.

These are the key factors that make it difficult to define fully the measured differences in consumption for the participant and comparison groups. 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 a mix of net and gross savings.

In the extreme, all household 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 and would not control for any natural energy efficient activity. This savings estimate would retain all of the freerider 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 as the participants and, as a result, are unlikely to exhibit any natural energy-efficient activity of the sort under evaluation. The use of customers who have participated in the same program in a recent year—or will participate in the near future (pipeline)—avoids most of the concerns related to self-selection bias. 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.

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. As a result, a comparison group created from previous and future participants may be as similar to current-year participants as is possible outside of an

RCT. Thus, the use of such a comparison group is likely to produce a gross estimate of savings that is unbiased due to self-selection.

4.2 Recommendations by Program Characteristics

The billing 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.

Table 1. Program Characteristics, Comparison Group Specifications, and Billing Analysis Structure and Interpretation

Program Condition	Billing Analysis Form	Comparison Group	Gross or Net Savings	Unknown Biases
1. Randomized Controlled Trial, Experimental Design	Two-Stage or Pooled	Randomly Selected Control Group	Net	Spillover, if it exists
2. Stable Program & Target Population Over Multiple Years	Two-Stage	Prior and Future Participants	Gross	Minimal
3. Participation staggered over at least one full year	Pooled	None: Pooled specification with Participants only	Gross	Minimal
4. Not randomized, not stable over multiple years, participants similar to general eligible population, nonparticipant spillover minimal	Two-Stage or Pooled	Matched comparison group	Likely between gross and net	Self-selection[7] and Spillover
5. Not randomized, not stable over multiple years, participants unlike general eligible population, nonparticipant spillover minimal	Two-Stage or Pooled	General Eligible Nonparticipants	Likely between gross and net	Self-selection and Spillover

Table rows 1, 2, and 3 provide at least one feasible approach for any whole-building retrofit program. Experimental design is still somewhat rare, but for many of the reasons discussed in this document, it is becoming more-widely used. A stable program makes possible the opportunity to obtain an unbiased estimate of savings using the Two-Stage approach.

Most other programs can be evaluated using the pooled approach. Rows 4 and 5 of the table list two relatively common approaches in the industry. These approaches produce an estimate that is a mix of net and gross savings. If this approach is used, then the result must be considered a conservative gross savings estimate with a known downward bias, to the extent freeriders still exist in the comparison group population. A separate freeridership analysis is required (for example, self-reported) to adjust all of these gross savings estimates to net savings estimates.

[7] The matched comparison should mitigate some self-selection to the extent that it is correlated with relative pre-period consumption, and this is an improvement over a non-matched, general population comparison group.

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

4.3 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 before the start of the current program year,[8] 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

4.4 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 utilizes 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.

8 It is counter-intuitive 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, similar in all ways to post-program participants. The difference-in-difference structure relies on an additive period-to-period change factor that works equally well with past or future participants.

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[9], 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; or
- 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

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

the analyst has the freedom to create these comparison group pre- and post- data periods using exactly the same distribution as the current year participant dates.

3.4 Basic Data Preparation

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

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 is 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.

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. (This activity is detailed in the section titled, "Data Requirements and Collection Methods.")

10. ***Calculate average daily degree days*** for the exact dates of each bill interval in the consumption data.

4.5 The Two-Stage Approach

4.5.1 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 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.[10] (The theory regarding the underlying structure is discussed in materials for and articles about the software.[11]) Stage 1 of the analysis can be conducted using PRISM or other statistical software.

4.5.1.1 Step 1.Fit the Basic Stage 1 Model

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

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

where

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(τ_H) during meter read interval m, based on daily average temperatures over those dates;

---

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

[11] Energy and Buildings: Special Issue devoted to Measuring Energy Savings: The Scorekeeping Approach. Margaret F. Fels, ed. Volume 9 Numbers 1&2, February/May 1986.

C_m = Specifically, $C_m(\tau_C)$, average daily cooling degree-days at the base temperature(τ_C) during meter read interval m, based on daily average temperatures over those dates;

μ = Average daily baseload consumption estimated by the regression;

β_H, β_C = Heating and cooling coefficients estimated by the regression;

ε_m = Regression residual.

4.5.1.2 Stage 1 Model Selection

4.5.1.2.1 Fixed Versus Variable Degree-Day Base

In the simplest form of this model, the degree-day base temperatures τ_H and τ_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 the savings estimation uses NAC only ***and*** the decomposition of usage into heating, cooling, and base components is not of interest. When 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 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.

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, adjusted R^2, AIC or BIC[12]) from among all of these models.

---

[12] 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, τ_H and τ_C across an appropriate range. This approach includes a specification in which one or both of the weather parameters are removed.

4.5.1.2.2 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 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 me estimated for each of these options: a heating-cooling model (HC), heating only (HO), cooling only (CO), and no degree-day terms (mean value).

4.5.1.2.3 Degree-Days and Setpoints

If degree-days are allowed to vary,

- The estimated heating degree-day base τ_H will approximate the highest average daily outdoor temperature at which the heating system is needed for the day, and
- The estimated cooling degree-day base τ_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 setpoint and building dynamics, such as insulation, and internal and solar heat gains.

The average thermostat setpoints 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 (from55^{o}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.[13] (Note that the cooling degree-day base must always be higher than the heating degree-day base.)

---

[13] 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, H0 or C0), select as the final degree-day bases the values of τ_H, and τ_Ct hat give the highest R^2, along with the coefficients μ, β_H, β_C estimated at those bases. Models with negative parameter estimates should be removed from consideration, although they rarely survive the optimal model selection process.

4.5.1.3 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 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^oF for heating and 70^oF 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[14]:

- If the heating and cooling coefficients in the HC model have p-values[15]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.
    - 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.***

---

[14] Adjusted R2, AIC or BIC are also used.

[15] A measure of statistical significance.

4.5.2 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 μ, $\beta_{H,}$ and β_C with the annual normal-year or typical meteorological year (TMY)[16] degree-days H_0 and C_0 calculated at the site-specific degree-day base(s), τ_H and τ_H. 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.

$$NAC = \mu * 365 + \beta_H H_0 + \beta_C C_0$$

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.

4.5.3 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.

4.6 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.

4.6.1 Recommended Forms of Stage-Two Regression

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 since they have different central tendencies and variances.

4.6.1.1 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.

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

where

ΔNAC_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

---

[16] Discussed in Section 4, Measurement and Verification Plan.

β, γ = coefficients determined by the regression

ε_j = regression residual.

From the fitted equation:

- The estimated coefficient γ is the estimate of mean savings.
- The estimated coefficient β is the estimate of mean change or trend unrelated to the program.

The coefficient β corresponds to the average change among the comparison group, while the coefficient γ 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.

4.6.1.2 Form B. Multiple Regression With Program Dummy Variables

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

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

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 β_0 serves as an intercept term.

β_q, γ_k = coefficients determined by the regression

From the estimated equation:

- The estimated coefficient γ_k is the estimate of mean savings per participant who received measure group k.
- The coefficient β_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; or

- With multiple dummy variables I_k and a single characteristics variable x (other than the intercept); or
- 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 a low-income program, these variables may be obtained from tracking data collected the same way across the program years.

Form C. Statistically Adjusted Engineering (SAE) 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.

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

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

β_q, ρ_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 β_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, or
- With multiple measure groups k and a single characteristics variable x (other than the intercept), or
- With 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, unless one of their associated coefficients is found to be statistically insignificant.

A simpler SAE form that omits the participation dummy variable has the nominal appeal of the coefficient ρ_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—or if it varies in ways that are not well correlated with actual savings—then the inclusion of the tracking estimate will not improve the fit. Thus, the dummy-variable version is preferred.

4.6.2 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, ***and***
- 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, ***and***
- 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, ***and***
- 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 billing 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.

5 Pooled Fixed-Effects Approach

The pooled approach addresses exogenous change without the inclusion of a separate comparison group. In this model, participants who received a measure installation during a certain time interval serve as a steady-state comparison for other participants in each other time interval.

Almost all observations include premises that are still in their pre-installation period ***and*** premises in that are in their post-installation period, so the effect of post- versus pre- is estimated to control for exogenous trends.

The basic structures of the site-level and the second-stage billing model are effectively combined in the pooled approach. All monthly participant 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) and average;
- Overall heating and cooling components; and
- A post-installation indicator variable capturing the change in the post-installation period.

5.1 Recommended Form of Pooled Regression

The recommended pooled model equation is as follows:

$$E_{im} = \mu_i + \phi_m + \Sigma_k \beta_{Hk} I_{kj} H_{jm} + \Sigma_k \gamma_k I_{kj} P_m + \Sigma_k \gamma_{Hk} I_{kj} H_{jm} P_m +$$

$$\Sigma_{kq} \beta_{Hkq} I_{kj} H_{jm} x_{qj} + \Sigma_{kq} \gamma_{kq} I_{kj} x_{qj} P_m + \Sigma_{kq} \gamma_{Hkq} I_{kj} H_{jm} x_{qj} P_m + \varepsilon_{im}$$

Where all variables have already been defined except for these:

μ_i	=	Unique intercept for each participant i,
ϕ_m	=	0/1 Indicator for each time interval *m*, time series component that track systematic change over time
P_m	=	0/1 Indicator variable for the post-installation period.

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

The parameter interactions that include the variable P_m capture the savings in the post-installation period. The inclusion of the read interval fixed-effects controls for exogenous factors specific to each month, and to first order eliminates the correlation across customers ij of residuals, ε_{im}, for a given month m.

If there is any intent to use the heating or cooling components of the model separately, the model should be fit across a range of degree-day base combinations. The highest R^2 is used to determine the optimal degree-day base combination[17].

From the fitted equation:

- The mean program savings must be calculated using the coefficients on all of the post-period dummy variable components, annual normal or TMY heating, and/or cooling degree-days for participants with measure k ***and*** the mean household characteristics (square footage, etc.) for households with measure k. That is, the estimated mean program savings for measure group k is

  $$S_k = \gamma_{k*}365.25 + \gamma_{Hk}H_{0k} + \Sigma_q \gamma_{kq}x_{qk_} + \Sigma_q \gamma_{Hkq}H_{0k}x_{qk_}$$

- WhereH_{0k} is normalized or TMY degree-days at the appropriate base for the subset of households with measure k, x_{qk} is the mean value of characteristics variable x_q for customers who received measure k.
- The coefficient ϕ_m is a monthly estimate of mean change or trend unrelated to the program. Because of the fixed-effects structure, these estimates represent the delta from the month or months left out of the model. That is, they are not mean zero and must be must be included if pre-treatment consumption is to be calculated.

In general, the increased complexity of the pooled approach requires additional care by the evaluator. The estimates of savings and consumption developed from any model must be carefully constructed and vetted against raw data. Developing a parallel two-stage model as a point of comparison for pooled model quality control should be considered.

5.1.1 Choice of Pooled Form

The pooled approach is recommended if:

- There is not a valid nonparticipant comparison group, or
- The goal is to measure an average savings effect over multiple program years.

In addition, the pooled approach requires both of the following:

- ***A balance of participant installation intervals across at least three billing intervals,*** preferably more. Having a balanced participation across three intervals would ensure that two-thirds of the participants provide a steady-state comparison during each interval of change. In the extreme, with only a single start date (as with a program that starts mailing comparative usage reports to homes at the same time), the model

---

[17] Note that the pooled model estimates average the heating and cooling degree day bases and average that slopes that are meant to represent the average across all homes in the model (or defined by interaction effects). This averaging can work well in many cases, but it can be difficult to determine when it may not work well. Therefore, if specific heating or cooling load components are of interest, the two-stage approach, which allows for house-specific degree day bases and heating/cooling slopes, may be a better choice.

fails to control for exogenous change across the change point. This explains the more stringent requirement for these programs of a randomly assigned experimental design.

- ***A balance of data between pre- and post-installation periods with respect to the number of data points per household and the seasonal coverage.*** Similar seasonal coverage in the pre- and post-installation is particularly important if measure savings are temperature sensitive. For gas heat modeling, the model should include at least one full winter in both the pre- and post- periods *and* some non-heating months. A full year of pre- and post-installation data removes concerns regarding imbalanced data.

The recommended specification includes the characteristics variables (x_j) for each house because of the importance of these factors:

- Having additional data to inform the overall average heating and cooling trends, and
- The changes in those trends due to the program.

In particular, it is useful to include a consistent square-footage variable. These characteristics data help compensate for the pooled approach's inherent lack of flexibility with respect to heating and cooling dynamics, as compared to the site-level model approach.

6 Measurement and Verification Plan

6.1 IPMVP Option C

The recommended IPMVP method is Option C (Whole Facility), which was designed in part to address evaluation conditions that occur with a whole-house retrofit program. The key reasons for using this method are these:

- The goal of the program is improvement of whole-house performance;
- Because multiple different measures are installed, the individual savings of each cannot be easily isolated because of interactive effects; and
- The expected savings are large enough to be discernible over natural variation in the consumption data, at least across the aggregate of program participants.

Major non-program changes in energy consumption are either not expected or will be adequately controlled for in the analysis.

6.2 Verification Process

This does not apply for whole house retrofit savings based on billing analysis.

6.3 Data Requirements and Collection Methods

A billing analysis requires data from multiple sources:

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

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

6.3.1 Billing Data

The consumption data used in a billing analysis are generally stored as part of the utility billing system. Since 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 billing 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 billing regression.

The remaining requested variables serve one of three purposes:

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

6.3.1.1 Billing 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 99th 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; however, do this 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 billing 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).

6.3.2 Weather Data

Weather data are used in the billing 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[18].

Use either primary 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 so as to produce evaluation results that are consistent with other studies within the utility. Many utilities are choosing to use normals constructed from fewer than 30 years, as are the standard NOAA norms.

A billing 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 that the same site. 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.

6.3.2.1 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 concern regarding a weather data source, consider using a more distance 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.

6.3.3 Tracking Data

The program tracking data provide the participant population, the pre- and post-installation time periods, and the number and type of measures for which savings are claimed. Frequently, the

---

[18] The National Oceanic and Atmospheric Administration (NOAA) produces 30-year normal weather series composed of average temperature for each hour over the time period. These normals are updated every decade. National Renewable Energy Laboratory produces typical meteorological year (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.

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 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 year may be used to define a control group for a Two-Stage analysis.

6.4 Analysis Dataset

Using the account numbers in the two datasets, the final analysis dataset combines the tracking data and the billing 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 of the unique read intervals, some evaluators resort to calendarized data.[19] Except in special cases, calendarization should not be

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

used for this kind of analysis because it undermines the direct matching between consumption and degree-days that is the basis of billing analysis. Multiple meter and multifamily analyses are examples of situation where calendarization may be the only way to aggregate data series on difference schedules.

6.4.1 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 billing 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 billing 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.

    - 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 billing 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.

7 Sample Design

Sample design is generally not required for whole-house retrofit billing analyses because this type of evaluation is performed on the full, relevant program population.

7.1 Program Evaluation Elements: Considerations for Other Program Types and Conditions

The methods described above are used in whole-building program evaluation for an ongoing, stable residential program. Similar methods can be used for the following: (a) other whole-premise programs for the residential population; (b) whole-premise programs for small commercial populations; and (c) with modification, for new construction.[20] Whole-premise billing analysis is also used for other types of programs, such as single-measure rebate programs and recycling programs.[21] In this section, we discuss the alternative comparison group specification to use in these situations.

7.2 Alternative Comparison Group Specifications

In some cases, it is not practical to use past or future participants as a comparison group, or to conduct a pooled billing 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, as discussed above.
- 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, a two-stage 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.

Matching is accomplished by: (1) Determining the mix in the participant population, and (2) selecting a stratified nonparticipant sample with the corresponding mix from those customers

---

[20] Discussed in a separate chapter.

[21] See Chapter on Furnaces and Boilers

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); and
- Energy end uses.

8 References and Resources

ASHRAE recommends that the following ASHRAE documents be listed as either a reference or resource document.

- Performance Measurement Protocols for Commercial Buildings, ASHRAE, 2010
- Research Project 1050 Development of a Toolkit for Calculating Linear, Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models, ASHRAE Research Report, 2004.
- ASHRAE Guideline 14-2002 Measurement of Energy and Demand Savings, ASHRAE, 2002.
- ASHRAE Guideline 14-2002R (Revision of Guideline 14, currently in process, publication date TBD).