Comparison of Bayesian Billing Analysis to Pooled Fixed Effects and Variable-Base Degree-Day

Benjamin Hannas, Ecotope, Inc., Seattle, WA
Michael Logsdon, Ecotope, Inc., Seattle, WA

ABSTRACT

Two popular billing analysis techniques are pooled fixed effects (PFE) and two-stage variable-base degree-day (VBDD) analysis. These frequentist methods are well documented and widely used, but, like all models, each has limitations. Frequentist billing analysis methods lack good estimates of error and the results can be biased, particularly for smaller data sets. Further, advances in computational power have allowed other methods to become more feasible. A Bayesian inference model is developed to overcome some of the limitations of the frequentist methods. The Bayesian approach provides summary statistics from the building-level up to the program-level, is consistent with respect to sample size, and provides a reliable estimate of the standard error. Bayesian analysis also uses more of the buildings in the final analysis compared to PFE and VBDD. VBDD loses buildings because of building-level model misspecification and other issues. A modification of VBDD is also presented using penalized regression as an attempt to avoid losing as many buildings in the final analysis. Cases studies are used to show effects of models on individual buildings. Running multiple models for evaluations is useful for comparison, and including the Bayesian approach as one of the models can overcome some of the limitations of PFE and VBDD approaches.

Introduction

Two popular billing analysis techniques are pooled fixed effects and two-stage variable-base degree-day analysis. These frequentist methods are well documented and widely used (examples include Berry 2003, Pigg 2002, Wang 2006), but each has logistical and analytical advantages and disadvantages (Berger 2013). In addition to the discussion in the Berger paper, frequentist billing analysis methods lack good estimates of error and the results can be biased, particularly for smaller data sets. Further, advances in computational power have allowed other methods to become more feasible. This paper is a review of typical billing analysis techniques, an exploration of adjustments to those techniques, and a presentation of a Bayesian inference framework. The philosophical difference between frequentist and Bayesian approaches, as written in Wakefield 2013,

> *"Central to the philosophy of each approach is the interpretation of probability that is taken. In the frequentist approach, as the name suggests, probabilities are viewed as limiting frequencies under infinite, hypothetical replications of the situation under consideration... By contrast, in the Bayesian approach...probabilities are viewed as subjective, and are interpreted conditional on the available information."*

The context of the discussion is on a whole-building residential evaluation across a large number of existing buildings. This longitudinal data set includes data trends within each building and trends across the population. If we use simple fixed effects regression in the analysis, then the two extreme options would be to analyze all data in a single model as if they were independent data points or analyze each building individually as if trends across buildings were independent of each other (Wakefield 2013). In the context of the models reviewed in this paper, the single model is pooled fixed effects and

the individual models are variable-base degree-day; a Bayesian approach is contrasted with these models.

In reality, for program evaluation we expect there to be a common trend across all buildings, but that bills for each building would be more similar to other bills from that building than to the general population. One solution for this mixed effects framework is Bayesian inference modeling—mixed effects being the combination of fixed effects across the population and random effects for each building.

The paper is organized into the following sections:

- Definition of Input Data
- Model Specifications
- Comparison of Model Specifications
- Case Study Model Comparisons
- Conclusions

Definition of Input Data

All of the models reviewed in this paper are regression models with energy consumption data as the dependent variable. Consumption data are typically monthly or bi-monthly (referenced as monthly within this paper). Independent variables can vary based on the type of measures installed at the buildings and/or the level of detail available for the buildings, but will at least include outdoor dry-bulb temperature, which is the most important regressor variable for typical buildings (ASHRAE 2013). For pre/post analysis, a variable indicating whether bills are pre- or post-installation of the measure is included (typically using a 0 or 1 as the indicator, or -1 and 1). If the installation includes a mix of measures, information about which measures are installed at each building can be useful. Other independent variables may also be included in the data set, such as occupancy, building vintage, building size, additional weather parameters, etc. The final data set is in a vertical format, with each observation (each line of the data) representing one utility reading at one building, along with the regressor and indicator variables mentioned above.

Ideal and Non-Ideal Data

The ideal data set has:

1. At least twelve months pre-data and twelve months post-data for every building,
2. Each bill is for approximately the same amount of time (no off-cycle reads),
3. A parallel comparison group with the same characteristics as the installed measure group,
4. No data issues (zero reads, extreme data, missing data, estimated reads, adjustment reads, overlapping read intervals), and
5. Buildings all have the same clear relationship with outdoor temperature (heating without cooling, for instance).

So what can go wrong in the real world? All of the above, usually. This is because a utility's need for regular meter reading is typically not as great as the needs of M&V (EVO 2012). In the case of data issues or not having enough bills in the pre and post time periods, buildings are usually dropped out of the analysis during the data-cleaning phase. More buildings may be dropped during the analysis phase if it is obvious the model does not properly describe the building (low R^2, for instance). Each of the models described below has different sensitivities to poor data, and it is important to acknowledge the deficiencies when performing the analysis since these issues will always be present in real-world data.

One final note about the input data: a good analysis will use multiple models to verify conclusions. An ideal input data set should allow the same data to be run through multiple models with minimal effort. One way to do this is to maintain the full original data, but add indicator variables based

on the needs of each model. For instance, a data set may have several logical indicator variables for whether a particular bill from a particular building should be included in the pooled fixed effects model, the variable-base degree-day model, and Bayesian inference model. Having this available in a single data set will help ensure the same base data are used in each analysis, as well as provide the basis for a summary table comparing the input data sets between the models.

Model Specifications

The methods reviewed in this paper are pooled fixed effects, variable-base degree-day, and Bayesian inference modeling. Other methods exist, as well as variations on these methods, but pooled fixed effects and variable-base degree-day models are commonly used and have extensive documentation (Agnew 2013, ASHRAE 2002, ASHRAE 2013, Fels 1986, EVO 2012). Bayesian inference methods are based on eighteenth century statistical concepts that gained popularity with early computing in the mid-twentieth century (Feinberg 2003). Increases in computational power have made these methods more accessible.

Change-point analysis is another common technique and is a variable-base analysis approach using average monthly temperature rather than degree-days (i.e., average daily temperature). Computationally, change-point analysis is the same as variable-base degree-day after the aggregation of weather data, but the use of average monthly temperature rather than average daily temperature makes the change-point analysis less connected to the physical model of the building compared to variable-base degree-day. This can be seen in months with large temperature swings where the degree-day approach may indicate both heating and cooling used in a month, but the change-point analysis only has the ability to select heating, cooling, or none, but not a combination of uses.

Pooled Fixed Effects

In the pooled fixed effects model, all data are included in a single model. The dependent variable is the energy use per billing period. Data include all buildings and observations where there are at least six months each of pre and post billing data after normal bill cleanup tasks (removing data issues from item 4 of the ideal data set list above, making sure weather and characteristics data are included, etc.). However, all seasons should be represented in both the pre and post periods when looking across all buildings in the data set (Agnew 2013).

The pooled fixed effects model assumes all observations are independent samples and generates an average curve of the data, which is used to find the average savings of the measure. However, because we expect the bills of a particular building to be similar to other bills for that building, our residuals are likely to be correlated by building and not be independent. Thus, our standard calculations for standard errors and uncertainty are likely incorrect (Wakefield 2013). Consequently, final estimates of error bounds, a useful piece of information for utility evaluators, can be unreliable.

Another issue with the pooled fixed effects model is the weather-dependent term (usually degree-days or average temperature). If we model each building individually (see variable-base degree-day below), we see each building has a different interaction with the weather data. Thermostat set-point, internal gains, insulation levels, and infiltration can affect the balance point temperature where conditioning needs change. The pooled fixed effects model can look at a range of balance points, but the balance point must be the same across the entire model. In general, the pooled fixed effects model does not have a direct, causal link to the physical reality of heating or cooling a building. The solution from the pooled fixed effects model is just a correlation to parameters, and the coefficients of the model can depend on parameter selection.

Variable-Base Degree-Day

The variable-base degree-day approach creates a physically-based heat-loss model for each building in each of the pre and post periods (Fels 1986). The approach involves two fixed effects regression routines, also referred to as the two-stage approach (Agnew 2013). The first stage is a fixed effects regression within each building and pre/post period, followed by the stage two fixed effects regression on the aggregated normalized results with characteristic data (occupants, measures, etc.).

Data for the first stage include the weather data and consumption data. Each building in the first stage should include at least nine months each of pre and post data spanning at least one heating season, one cooling season, and one shoulder season after normal bill cleanup tasks (Agnew 2013). (For gas bills, six months including half of the winter and some summer months can suffice.) This is a stricter screening criterion than the pooled fixed effects model, meaning more buildings are lost in this step compared to the similar step in the pooled fixed effects data processing.

Stage one of the variable-base degree-day approach generates a physically-based heat-loss model for both the pre and post measure installation periods for each building. These models are then normalized to typical weather and the pre and post models are compared to estimate the savings on a per-building basis, but only if the models are acceptable. There will be a number of buildings where the pre and/or post models are not acceptable because:

- The fit is not good enough (using R^2 or similar metric),
- The balance point temperature boundary was reached (if we search between 40 and 70 degrees and return a balance point of 70, we should not trust the result),
- The slope is outside of what is considered "normal" (such as a negative slope, suggesting heating decreases with decreasing outdoor temperature, or no slope, or extremely high slope), or
- Other issues (reviewing the relationship reveals other odd behavior, like vacations, etc.).

Model misspecification is a common issue of poor fits. For instance, specifying a heating model, but heating is overshadowed by baseload, or non-utility fuel is used and there is no electric or gas heating signature, or a heating-only model is specified for electric use but there is a lot of cooling use (causing the baseload estimate to be artificially high). Some of these issues can be avoided by using a more sophisticated algorithm. In general, though, variable-base degree-day will have more buildings removed in the final analysis than the pooled fixed effects model. In the pooled fixed effects model, all of these buildings can be included in the model without knowledge or obvious indication of potential deficiencies contained therein.

Stage two of the variable-base degree-day approach is a regression model using the savings results of the individual model fits from stage one. These stage one savings are obtained from applying typical weather data to the pre and post stage one model parameters and calculating the difference in the results. The per-building savings from stage one is the dependent variable in the stage two regression, and the independent variables of stage two are the building characteristics (typically the same characteristics as the pooled fixed effects model, without the weather since stage one already accounted for that effect).

Standard errors can be calculated for both stage one and stage two results (Fels 1986). However, the stage two standard errors do not account for the errors found from stage one—instead, the stage two standard error is calculated only from the distribution of the savings point estimates found in stage one. Another issue with the stage two model is the loss of buildings during data cleaning and during the stage one analysis. This can generate known or unknown bias in the results. Important building characteristics should be compared between the dropped buildings and the buildings kept in the analysis to check for bias, although even with similar building characteristics discarding data on the basis of R^2 or other metrics has likely unknown implications for the results.

Adjustments to Variable-Base Degree-Day

Variable-base methods can suffer from poor model fits due to misspecification of the underlying model for a building or from a few outliers in the data. Improvements can be made to the traditional methods in hopes of reducing the number of buildings dropped in the analysis. Reducing the number of buildings removed will help prevent bias in the final savings estimate because the model will apply to a larger fraction of the population.

The proposed approach is to use penalized regression instead of standard regression (see Hastie 2013 for more information). Penalized regression can reduce the influence of outliers, creating a more realistic model. Penalized regression can also help with model specification, deciding whether a one-parameter, three-parameter, or five-parameter model best fits the data for a particular building. Traditional approaches to variable-base degree-day have fit a single model type to all buildings, and the models have been free to find a balance point across a wide range of temperatures.

In the Northwest, most buildings do not have a strong cooling signature when applying variable-base degree-day or change-point analysis on monthly data, so all buildings are run with a three-parameter model (Rushton 2014). When working with thousands of buildings, it is time-prohibitive to review each building to assign different models, so using a single model that fits a majority of buildings is useful.

A naïve solution to picking models on a per-building basis would be to try a one-, three-, and five-parameter model for each building and pick the best model. However, without further conditions on the model, the five-parameter model will always win because it has more freedom to adjust to the data. Penalized regression can prevent this model over-specification by penalizing the higher-order models. We can set the penalty so that most buildings still use the three-parameter model, but some buildings will fit a five-parameter model and other buildings a one-parameter model.

Figure 1 shows an example of how the penalized regression works for a single building. The model is dependent on the size of the penalty, with a zero penalty creating a five-parameter model (heating and cooling model) and a high penalty creating a one-parameter model (average energy use with no relation to temperature). For a data set in the Northwest, we would select a penalty so that most buildings would select the three-parameter model, with some buildings selecting the five-parameter model and some selecting the one-parameter model.

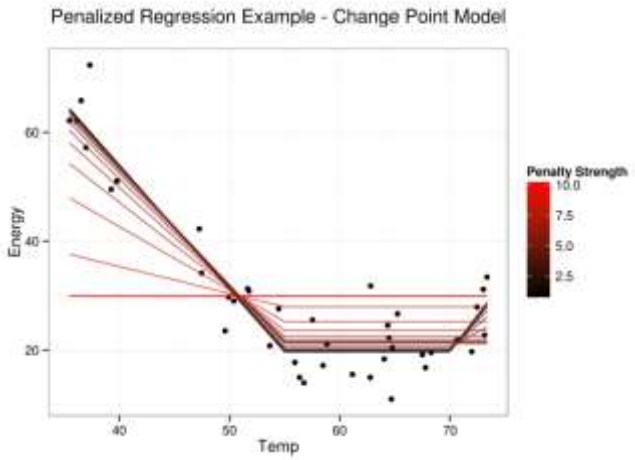

Figure 1. Penalized Regression Example for Frequentist Change-Point Model

The penalized regression tailors the model selection to the building-level data, but it also decreases the effect of odd data points on the model. Even after cleaning up the data, there will still be data points that are different from the expected behavior. If there are many of these data points, then we

consider them normal behavior, but if there are only a few odd data points then we might want to put less emphasis on those points. Figure 2 shows two examples of the penalized regression against the standard routine (labeled as Least Squares). In the right graph, a single data point causes the change-point to drop down to under 30°F and increases the slope dramatically. Applying a penalized regression model places the change-point and slope parameters into more reasonable values. In the left graph, there are a few high data points, but one extremely high point. The standard least squares approach again puts a lot of emphasis on the one very high data point. In the penalized approach, the one high data point has less effect, but the other few high data points still pull the fit higher since we believe there are systematic behaviors that will likely continue into the future, so we want to account for this extra energy.

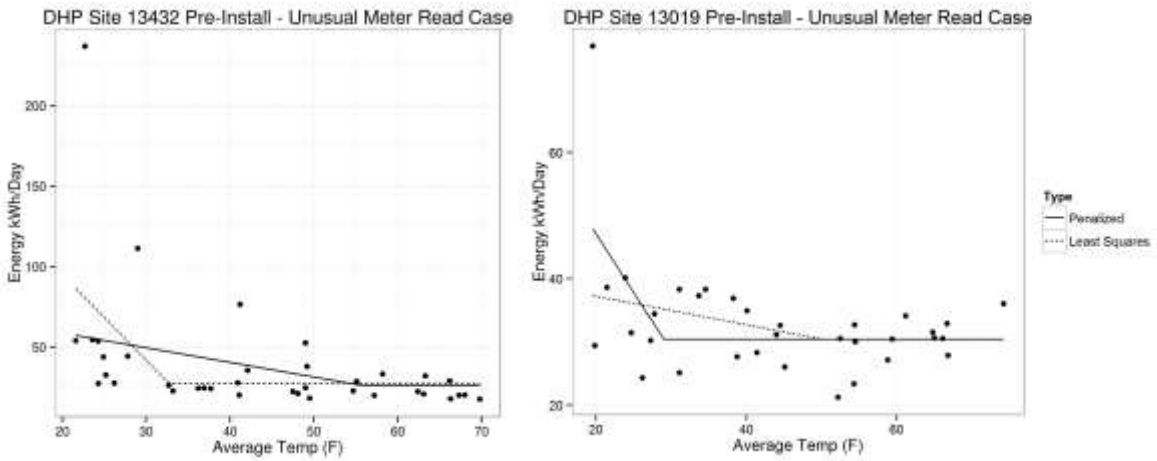

Figure 2. Two Examples of Comparing Penalized and Non-Penalized Frequentist Change-Point Models

The effect of the penalized regression is to place the change-point and the slope parameters into a more reasonable location. The goal is to generate models for more of the buildings that we would otherwise throw out of the final analysis because of poor model specification. With fewer buildings thrown out, we would expect less bias in the final result. This penalized approach is a forced process and the magnitude of the penalty can be slightly arbitrary, but it does create less extreme parameters and results in less dropped buildings.

Bayesian Inference

Bayesian inference methods are a mixed effects approach where all the data can be used in a single model, but also where data are recognized as dependent by building. Mixed effects models are a combination of trends across the population (fixed effects) and trends within each building (random effects).

The Bayesian approach has the least restrictive data requirements of the three models in this paper. Standard data cleanup should be exercised, but we do not need to filter for a certain number of pre and post data points per building. Using more of the original data will help alleviate building selection bias in the results.

One of the strengths of the Bayesian approach is the ability to borrow strength from the full model where necessary. This means if a building does not have a strong fit, then the fit across the population will assist in generating a model for that building. For comparison, in the case of variable-base degree-day, we would likely drop a building that does not have a good fit. If a building does have a good fit in the Bayesian approach, then the variable-base degree-day solution will likely agree with the Bayesian solution for that building.

The Bayesian approach is based in probabilities. In addition to defining the model to use, we also define probabilities for each term in the model. These initial probabilities are called the priors of the model and represent our best estimate of the parameters and their distributions before analyzing the data. The Bayesian routine essentially evaluates a complicated mathematical integral, combining the priors with the observed data to generate what are called posterior distributions. As this type of integration in general cannot be performed analytically, it is done from among a variety of numerical methods. It is important to note that, as more and more data are collected, the influence of the prior wanes and the estimated posterior distributions become driven more and more by the observed data. Reliance on subjective prior information has been one criticism of Bayesian inference, although in the scenario of energy use in buildings it may prove a strength, since we know a priori from physics the general configuration for how a building should use energy with respect to weather.

A benefit of this probabilistic approach is all outputs of the Bayesian model have both a point estimate and an estimate of the error. The reliable estimate of the standard error of the model is a big advantage over the previous models since we can now report both the measured savings and a credible interval of that savings.

The final posterior model is a very detailed model. If we have 1,000 buildings in the analysis, then the final model will have a parameter estimate and standard error for every fixed effects parameter as well as 1,000 parameter estimates for every random effects parameter. This allows savings estimates to be calculated on average for the population and for a specific building.

The main drawbacks of this approach are the model specification, implementation of the analysis, and possibly computation time, depending on the complexity and scope of the model and dataset in question. The model specification is much more complex than with the frequentist methods. The equation itself is more complicated, but we also need to specify distributions for all of the parameters. Implementation of this method requires a statistical software package for all but the simplest data sets. Our current implementation uses the Stan interface for R (Stan 2015), though the most recent Stata release also includes extensive tools for Bayesian analysis. RStan uses a sampling based approach to evaluating Bayesian models, rather than a direct or approximate solution to the integration. The unusual, piecewise nature of the mean model (heating, cooling, deadband) makes the implementation of sampling-based approaches much simpler to implement, although such methods can be computationally expensive.

Comparison of Model Specifications

Consider the following theoretical program evaluation. We have an electric heating retrofit across three heating zones in a northern climate. We expect the bills to have a strong heating signature, but only a low percentage of buildings will have a cooling signature. Our goal is to provide a savings estimate for the whole program and by climate. For simplicity with this example, we are not including a discussion of a comparison group.

The first step is to review the data for obvious issues—zero reads, extreme data, missing data, estimated reads, adjustment reads, overlapping read intervals, etc. Also, check that each building has been assigned weather data and a climate zone. It is useful for record keeping not to remove any data, but to add indicator variables to flag whether an observation should be included in the final analysis for each analysis method. Observations with major data issues should be removed for all methods. Next, add an indicator variable for each analysis method, since each method has a different threshold for keeping buildings in the analysis:

- Pooled fixed effects requires at least six pre and six post bills per building.
- Variable-base methods require a mix of winter, shoulder, and summer months in both the pre and post period per building.

- Bayesian analysis can proceed with limited further reduction in the data set.

We now specify the models. The dependent variable is the energy consumption per month per building. The independent variables include the temperatures per building per month and the climate zone per building.

Frequentist Pooled Fixed Effects

The pooled fixed effects model can be specified as:

$$E_{ij} = \alpha_1 + \alpha_2 T_{ij} + \alpha_3 P_{ij} + \alpha_4 C_i + \alpha_5 T_{ij} P_{ij} + \alpha_6 T_{ij} C_i + \alpha_7 P_{ij} C_i + \alpha_8 T_{ij} P_{ij} C_i + \varepsilon_{ij}, \quad (1)$$

with

$$T_{ij} = \left(\tau - t_{ij}\right)^+, \quad (2)$$

where t_{ij} is the temperature for building i and time period j, and T_{ij} is most commonly calculated as degree-days. The parameter τ is the base temperature across all buildings. The + sign refers to using positive values and setting negative values to 0. The model can be run for a number of different base temperatures to find the best fit, but the base temperature must be the same for all buildings. Other terms in the model are:

- α_* are the coefficients produced by the regression,
- P_{ij} is an indicator variable, defined as 0 for pre installation and 1 for post installation for building i in month j (could also be -1 and 1),
- C_i is the climate zone for building i, and
- ε_{ij} is the residual for building i in month j.

Savings from the pooled fixed effects model are calculated from the resulting fitted equation, but using average normal weather data in place of the weather data during the study period. Adjusting the climate parameter will obtain savings by climate zone, but each climate zone will need its own average normal weather. The fitted equation model is run for both the pre and post value for P_{ij}, and the difference in the two models outputs is the average savings.

Frequentist Variable-Base Degree-Day

The variable-base degree-day approach is based on a heat loss model of the building, which for the heating-only model is:

$$E = \gamma + \frac{\sum UA + V\rho C_p}{\theta}(T_{in} - T_{out})^+ - \frac{Q}{\theta}, \quad (3)$$

where

- E is the energy consumption,
- γ is the baseload consumption,
- The UA term is the sum of the conductive heat losses of the building,
- $V\rho C_p$ is the infiltration loss of the building,
- Q is the internal gains of the building,
- θ is the heating system efficiency, and
- $T_{in} - T_{out}$ is the difference in temperature from inside to outside (non-negative).

With rearrangement and simplification of the terms and addition of an error term and indices, this reduces to the variable-base degree-day model:

$$E_{ijk} = \beta_{1ik} + \beta_{2ik}\left(\tau_{ik} - t_{ijk}\right)^+ + \varepsilon_{ijk}. \quad (4)$$

Here, each building i in each pre/post period k has its own base temperature τ_{ik}. The base temperature τ_{ik} includes indoor temperature, heat loss of the building, and internal gains—see Appendix 1 of Fels 1986 for more details. The base temperature is used to calculate the daily difference in base-to-outdoor temperature and the sum of these values per month are the monthly degree-days used in the

model. The parameters from the best fit regression (β_*) are then used to calculate the pre and post normalized consumption, N_{ik},

$$N_{ik} = 365 \cdot \beta_{1ik} + \sum_m \beta_{2ik}(\tau_{ik} - t_{0m})^+, \quad (5)$$

where the temperature term is the total heating degree-days for the year. The savings for each building, denoted ΔN_i, is the difference between the normalized pre-consumption and the normalized post-consumption. The second stage regression is then:

$$\Delta N_i = \beta_3 + \beta_4 C_i + \varepsilon_i. \quad (6)$$

Results from the second stage regression are then used to find the average savings, both overall and by climate.

Bayesian Approach to Variable-Base Analysis

The model specification for the Bayesian change-point analysis is a combination of the previous two model concepts, but also includes a detailed specification of the distributions. We start with a modified version of Equation 1,

$$E_{ij} = (\alpha_1 + \alpha_{1i}) + (\alpha_2 + \alpha_{2i})\hat{T}_{ij} + (\alpha_3 + \alpha_{3i})P_{ij} + (\alpha_4 + \alpha_{4i})C_i + (\alpha_5 + \alpha_{5i})\hat{T}_{ij}P_{ij} + (\alpha_6 + \alpha_{6i})\hat{T}_{ij}C_i + (\alpha_7 + \alpha_{7i})P_{ij}C_i + (\alpha_8 + \alpha_{8i})\hat{T}_{ij}P_{ij}C_i + \varepsilon_{ij}, \quad (7)$$

with

$$\hat{T}_{ij} = \left(\tau + \hat{\tau}_i - t_{ij}\right)^+. \quad (8)$$

Equation 7 is based on Equation 1, where we keep all of the fixed effects terms from the first equation and add the per-building random effects, denoted as the α_{*i} parameters. The temperature parameter also receives a random effects term, $\hat{\tau}_i$, where $\tau + \hat{\tau}_i$ is the building-dependent temperature term τ_{ik} seen in Equation 3.

The second part of the Bayesian model specification is setting the priors. We will assume all of the random effects have a normal distribution centered on 0 with some variance, σ^2, meaning there is an equal chance for any given building that the random effects parameter is above or below the fixed effects model parameter. We then go through all of the α_* parameters from the fixed effects part of the model and define our priors based on engineering judgement and experience. For instance, a good estimate for the average baseload is 30 kWh/day and baseload cannot be a negative value, so we assume a prior on α_1 to be a gamma distribution with parameters set so the mean of the distribution is 30.

Average savings for the program (overall and by climate) are found by using the normalized weather for each building to calculate the average savings per building using the posterior model. Because we have distribution estimates for all parameters in the model (as they relate to both population and building-specific effects), we can also produce a distribution estimate for average saving by climate zone and overall.

Case Study Model Comparisons

The following is a case study comparison of building-level results between the frequentist variable-base approach and the Bayesian variable-base approach. The frequentist and Bayesian variable-base routines allow a comparison of building-level models, which is useful in evaluating how well a model fits the data. The pooled fixed effects model only provides aggregate statistics across all buildings, so is not included in this building-level comparison. Figure 3 shows two building-level comparisons of the variable-base models. On the left, data are well behaved and the frequentist model matches the Bayesian model. On the right, data are less behaved. There is much more variation in the relationship between energy and temperature, and the data point in the bottom-right corner is a strong anchor point for the frequentist model. Because of this one point, the frequentist model is placing the change-point at 70°F, which would typically cause this site to be discarded in the case of a change-point

search bounded by 70°F. In the Bayesian model, this bottom-right data point is still used in the model, but because we have information about the rest of the population as well as a prior that requires substantial evidence to conclude a high change point, the Bayesian model is unlikely to place a change point at 70°F. Instead, the change-point is around 60°F, which is much more in line with our experience with heating base temperatures.

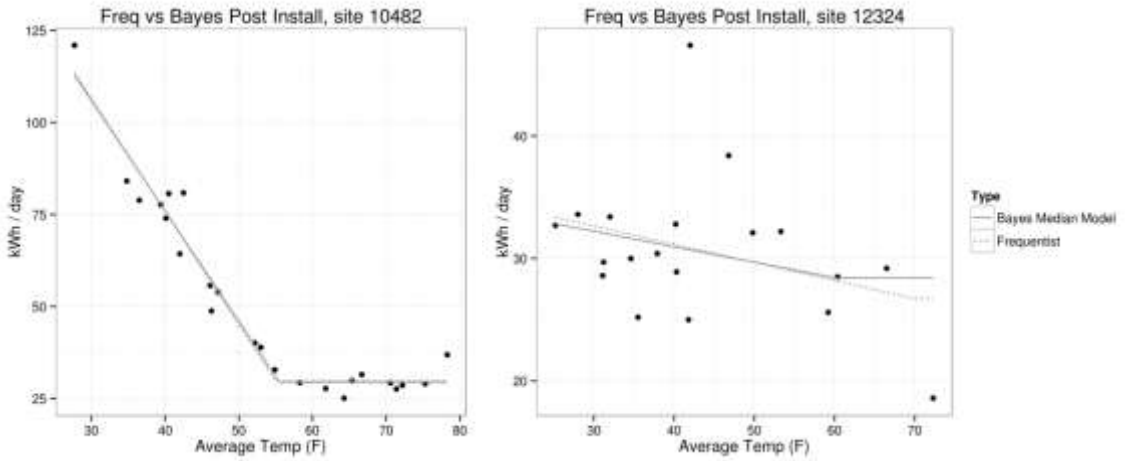

Figure 3. Comparison of Building-Level Change-Point Analysis for Well-Behaved (Left) and Less-Behaved (Right) Buildings

Program-level comparisons are the next step in comparing these analysis techniques. With program-level savings, results from all the methods above can be compared. The routines for doing this analysis are being finalized and two data sets have been selected for analysis. One of these data sets is an electric HVAC retrofit measure with monthly utility bills, which is similar to many pre/post evaluation data sets. The second data set is a detailed metering study where all HVAC equipment (as well as many other loads in the home) have detailed measured consumption. While this second data set is not a pre/post scenario, it does provide the basis for comparing measured heating and cooling energy use to the estimated heating and cooling energy use from the methods in this paper.

Conclusions

There are many advantages of using a Bayesian approach for residential utility bill analysis. All bills from all buildings can be included in a single model, but we can still recover building-level information from the result. The posterior estimates have reliable standard errors, so we can report both an estimate and an interval for the final savings. The model can also include characteristics, so we can summarize the final result by climate zone or measure type or other characteristics entered into the model. The drawbacks of the Bayesian approach are the upfront effort of specifying the model and the computing power necessary for running a large billing analysis. However, the benefits of the approach are much greater than these drawbacks.

Even with the advantages of the Bayesian approach in almost every aspect, it is still useful to run other types of models for verification of the results. If the data are organized well from the beginning, analyzing the data set using frequentist approaches is inexpensive if a Bayesian analysis will already be performed. Pooled fixed effects just requires a quick check of pre vs post billing data points and a model specification before being run, though the model specification will likely be related to the Bayesian model. For variable-base degree-day analysis, a similar check of pre vs post billing data and model selection for stage two is necessary, as well as whether the standard approach will be used or a penalized

regression approach. The penalized regression can increase the number of buildings included in the final analysis, which can reduce the chance of unknown bias in the final result.

A package of billing analysis tools is currently under development for the R language. Having a common set of tools for analyzing data will make quick work of running multiple models. The upfront work of collecting bills, cleaning the data set, and defining building characteristics will still be a huge task, but after a clean data set is available, running multiple models on the same data set using standard tools should allow for more consistent and clear results. The models defined in this paper have been added to the package and full testing using real and artificial data sets will begin later this year, which will allow comparison of savings both at the building level and at the program level across multiple models.

References

Agnew, K., M. Goldberg. 2013. "Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol." *The Uniform Methods Project*. NREL/SR-7A30-53827.

[ASHRAE] American Society of Heating, Refrigerating, and Air-Conditioning Engineers. 2002. "Measurement of Energy and Demand Savings." *ASHRAE Guideline 14*.

[ASHRAE] American Society of Heating, Refrigerating, and Air-Conditioning Engineers. 2013. "Energy Estimating and Modeling Methods." *ASHRAE Handbook—Fundamentals*. Chapter 19.

Berger, J., F. Ucar. 2013. "Comparison of Pooled and Household-Level Usage Impact Analysis." *International Energy Program Evaluation Conference*. Chicago.

Berry, L., M. Schweitzer. 2003. "Metaevaluation of National Weatherization Assistance Program Based on State Studies,1993–2002." Oak Ridge National Laboratory. ORNL/CON-488.

[EVO] Efficiency Evaluation Organization. 2012. "Concepts and Options for Determining Energy and Water Savings." *International Performance Measurement and Verification Protocol* 1: 25-28.

Feinberg, S. 2003. "When Did Bayesian Inference Become 'Bayesian'?" *Bayesian Analysis*. http://www.stat.cmu.edu/~fienberg/Statistics36-756/BA-Fienberg-8-31-04.pdf.

Fels, M. 1986. "PRISM: An Introduction." *Energy and Buildings* 1986 (9): 5-18.

Hastie, T., R. Tibshirani, J. Friedman. 2013. *The Elements of Statistical Learning*. Springer Press.

Pigg, S. 2002. "Energy Savings from the Wisconsin ENERGY STAR Homes Program." Energy Center of Wisconsin. Research Report 211-1.

Rushton, J., A. Hadley. 2014. "SEEM Calibration, Phase I: Adjustments to Reflect Total Heating Energy, 2014 Update." RTF Staff Technical Report. http://rtf.nwcouncil.org/measures/support/SEEM/Default.asp.

Stan website. 2015. http://mc-stan.org/.

Wakefield, J. 2013. *Bayesian and Frequentist Regression Methods*. New York: Springer Press.

Wang, Y., et al. 2006. “Energy, Economic and Environmental Impacts of the Delaware Low-Income Weatherization Assistance Program.”