Lawrence Berkeley National Laboratory

Evaluation of Methods to Assess the Uncertainty in Estimated Energy Savings

Samir Touzani[1]
Jessica Granderson[1]
David Jump[2]
Derrick Rebello[3]

[1]Lawrence Berkeley National Laboratory
[2]kW Engineering
[3]Quantum Energy Services & Technologies Inc.

Energy Technologies Area
June 2019

Please cite as:
Touzani, S., Granderson, J., Jump, D. and Rebello, D. (2019). Evaluation of methods to assess the uncertainty in estimated energy savings. Energy and Buildings, 193, pp.216-225.

Disclaimer:

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Evaluation of Methods to Assess the Uncertainty in Estimated Energy Savings

Samir Touzani[1], Jessica Granderson[1], David Jump[2] and Derrick Rebello[3]

[1]*Lawrence Berkeley National Laboratory, 1 Cyclotron Rd., Berkeley, CA 94720, USA*
[2] *kW Engineering*
[3] *Quantum Energy Services & Technologies, Inc.*

Abstract

In this work we present a methodology to evaluate the accuracy of methods used to quantify the uncertainty in estimated total energy savings. We focus on savings measurement and verification (M&V) approaches that use a baseline model to characterize energy use, and that forward-project the model for a counterfactual to determine avoided energy use. These approaches are common to the International Performance Measurement and Verification Protocol's (IPMVP's) Option C and Option B. This methodology can be used to evaluate the uncertainty in savings estimates that are due to model error. It has been applied to evaluate two uncertainty estimation methods, including the industry standard ASHRAE Guideline 14 approach. The evaluation used data from 69 commercial buildings and four different baseline models that span daily and hourly granularity, as well as linear and non-linear/non-parametric forms. The findings of this work indicate that the standard methods that are widely used by the M&V community for estimating the total savings uncertainty over the post-installation period tend to underestimate the uncertainty. The tendency to underestimate the uncertainty is stronger for hourly models than for daily models, due to stronger autocorrelation in model residuals at the hourly time scale.

Introduction

In 2012, commercial buildings in the US consumed nearly 7 quadrillion Btu in site energy [EIA 2016]. Utility demand side management programs are the primary vehicle to deliver energy efficiency in the US building stock, representing investments of $8B (CEE 2017). Energy service companies (ESCOs) also represent a large share of the efficiency market, with $5B in revenue as of 2014 (Stuart et al. 2016). In these industries, reliable measurement and verification (M&V) of energy savings is critical, as it serves as the foundation financial settlement.

To date, the most commonly applied M&V approaches have relied upon engineering calculations and stipulated or deemed estimates of savings. In these cases, savings are treated as point values with no explicit assessment of uncertainty. Although M&V references such as ASHRAE Guideline 14 (ASHRAE 2014) address savings uncertainty due to model error, in practice, uncertainty analysis is most often applied in utility program impact evaluations as a means of ensuring that sampling plans appropriately reflect the program population.

Two trends are driving increased interest in meter-based whole-building level savings estimation, and in the uncertainty associated with those estimates – particularly in utility program

applications. These savings approaches are referred to as Option C in the International Performance Measurement and Verification Protocol (EVO 2012). First is a push to look beyond lighting and equipment replacement to identify the next generation of measures and program designs that will continue to deliver savings as more traditional measures begin to saturate. Of particular promise are operational, behavioral and retrocommissioning measures, strategic energy management, and multi-measure whole-building retrofits. These strategies offer deeper savings potential, are difficult to deem or calculate, and often involve multiple interactive effects, making them well suited to a whole-building meter-based approach. Similarly, pay for performance program designs that incentivize customers or implementers based on achieved savings are also well aligned with facility-level meter based savings estimation approaches.

Second, energy analytics technologies and the availability of smart meter data have converged to bring data science to the building energy efficiency industry. New techniques for building load prediction are increasingly being applied to diagnostic and control problems (Afram and Janabi-Sharifi 2014, Mamidi et al. 2012, Najafi et al. 2012), with extensions to savings estimation. Today's analytics technologies and advanced energy modeling applications are moving beyond the linear and piecewise linear approaches that have been used in the buildings industry for decades (Kissock et al. 2002, Fels 1989). More complex machine learning solutions that are based on higher frequency smart meter data (hourly or 15-minute) are beginning to be explored for their promise in increasing predictive accuracy (Granderson et al. 2016, Ahmad et al. 2017, Araya et al. 2017, Touzani et al. 2018). In addition, higher frequency data affords new opportunity to link efficiency to grid considerations, by opening the door to time-dependent and location-resolved savings valuation.

Within the context of these trends, there is renewed interest in the ability to assess the uncertainty in a savings estimate that is due to model error. This uncertainty can be a useful risk management tool to ensure that whole-building level savings estimates are robust enough to use as the basis of financial settlement. It can also provide a means of assessing tradeoffs between depth of savings and model goodness of fit in cases where signal-to-noise concerns may bring facility-level savings measures into question. And in a general sense, the use of uncertainty due to model error may prove beneficial for verifying that proprietary algorithms provide acceptable results.

In the M&V context the focus is on total estimated energy savings over the post period, as opposed to the estimated savings at each *time step* over the post period; therefore, the uncertainty of the total energy savings is the uncertainty parameter of interest. The energy savings estimate provides a single point value that describes the performance of the implemented measure in the project, however the usage of this unique value raises a question of how accurate it is? The uncertainty can be seen as an interval of doubt surrounding the estimated value of energy savings - the true value of the savings is expected to be within this interval at some level of confidence.

ASHRAE Guideline 14 (ASHRAE 2014) provides analytical formulas (Reddy and Claridge 2010) to calculate the savings estimates uncertainty. Several variations of the ASHRAE Guideline 14 formulas that have been introduced in the literature, and a brief review of these alternative versions can be found in Koran 2017 and Koran et al. 2017. While the uncertainty quantification is well defined for the classical monthly linear model, there is no guidance for accurately quantifying the uncertainty when a high frequency or non-linear/non-parametrical model is used; similarly, it is unknown the extent to which the uncertainty formulations break down if applied to anything other than monthly linear models. Several savings uncertainty estimation methods have been introduced in the literature for non-linear/non-parametric baseline models, however these approaches have been developed to provide uncertainty estimates at each time step (over the post period) and not the uncertainty of the total energy savings (Subbarao and

Reddy 2011, Heo and Zavala 2012), or are purely qualitative (Walter et al. 2014). Accordingly, this work addresses the research question: *What is a methodology and key metrics required to evaluate algorithms that estimate savings uncertainty due to baseline model error?*

To answer this question, two approaches to estimating savings uncertainty are evaluated and compared: K-fold cross validation and the method presented in ASHRAE Guideline 14. The effectiveness of these approaches on different model types (high frequency, linear, and non-linear/non-parametric) was assessed using four different baseline models that span daily and hourly granularity, as well as linear and non-linear/non-parametric forms. The impact on the results of two different training periods is also analyzed - 12 months and 6 months. Note that concurrently to our work, Koran et al. 2017 proposed a study that aims to provide a comparison of different savings estimates uncertainty approaches: ASHRAE Guideline 14 formula, a revised version of the ASHRAE formula, an exact algebraic approach for ordinary least squares regression and a bootstrap approach. Their works differs from ours mainly on the fact that their analysis focused on comparing the different methods, whereas we are analyzing the accuracy of the uncertainty estimates.

2. Methodology

2.1 Baseline models

The baseline models used in whole building M&V are empirical models that relate energy usage to parameters such as outdoor air temperature, humidity, or building operating schedule. These models are developed using consumption data before an efficiency measure was implemented (i.e., pre period). The models are projected into the post period to estimate what the energy use would have been if the measure had not been implemented. The difference between the estimated and the metered energy consumption is taken as the avoided energy use or energy savings (Figure 1). Statistical/machine learning regression methods are a standard approach used for developing baseline models that aim to model the relationship between the response *y*, which is the pre-retrofit whole-building energy use and a set of independent variables (also known as explanatory variables) $\boldsymbol{x} = (x^{(1)}, \dots, x^{(d)})$, where *d* is the number of independent variables. For example, the input variables can be time of the week and the outdoor air temperature. Mathematically the regression problem can be represented for a given observation set $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\}$, as

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \qquad (1)$$

where $\boldsymbol{x}_i = (x^{(1)}, \dots, x^{(d)})$, $i = 1, \dots, n$ are d dimensional vectors of inputs variables, ε_i is independent Gaussian noise with mean 0 and variance σ_ε^2. Building a baseline model consists of approximating the function $f(\boldsymbol{x})$ given a set of observation $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\}$.

Figure 1. Actual and model-predicted energy data, overlaid with outside air temperature, for a 12-month pre-installation period (training period) and 12-month post-installation period (prediction period).

In recent years, several baseline energy modeling approaches that use interval meter data have been introduced in the academic literature and in the industry. These methods are based on traditional linear regression, nonlinear regression, and machine learning regression methods. In this study four different baseline models are used, and described in the following. These models were chosen in part, because they represent a cross-section of linear, non-linear/non-parametric, and daily and hourly frequency that allow a thorough investigation of conductions under which uncertainty estimates may break down.

Time-of-Week-and-Temperature model (TOWT)

The TOWT model (Mathieu et al. 2011) is a baseline model that includes time of the week (i.e., hour of the week), and piecewise-linear temperature response with fixed change points that were set to 45, 55, 65, 75 and 85˚F. In addition, different regression models are fitted for occupied and unoccupied periods of the day that were determined using the following procedure: a linear regression model is fitted using two independent variables that are defined using the outside air temperature as number of degrees below 50˚F for the first one and number of degrees above 65 ˚F for the second one. The time step is defined as occupied if most of residuals from this simple model where positive, which means that the building is using more energy than it was predicted, otherwise the time step is defined as unoccupied. The choice of the TOWT model was motivated by the fact that it has been shown in previous study (Granderson et al. 2016) to be highly accurate. In this work, we used the implementation of the TOWT model that is available within the RMV2.0 R package (Touzani and Granderson 2017). Note that this version of the TOWT model has a hyperparameter, which correspond to a weighting factor that gives more statistical weight to days that are nearby to the day being predicted (the default value has been used, i.e., 15).

Bayesian additive regression trees (BART)

Using the BART algorithm as the regression method, this model characterizes energy consumption using the following independent variables: the time of the week (i.e., hour of the week), hourly outdoor air temperature and a dummy variable that is equal to 1 if the considered day is a holiday and 0 if not. BART is a Bayesian nonparametric regression model, which can be used to estimate the energy consumption as the sum of several regression trees (i.e., decision trees), and can be seen as a Bayesian version of ensemble methods such as random forest and gradient boosting machine. BART differs from other ensembles of regression tree algorithms in

that it is fully Bayesian model-based, and as such it consists of several prior distributions for all unknown parameters that characterize the regression trees. The posterior is computed using Markov Chain Monte Carlo algorithm (Kapelner and Bleich 2016). One of the most important advantages of BART over more traditional ensemble regression trees algorithms (i.e., random forest and gradient boosting machine) is that it is less sensitive to the choice of the hyper-parameters that define the priors, making it easier to use by a non-expert in machine learning. Chipman et al. (2010) provided default settings for these hyper-parameters that simultaneously produce good fit and avoids over-fitting the training data. In this work we have used the proposed default values (Chipman et al. 2010) to build the baseline models, however it is important to note that it is possible to increase the prediction accuracy of the models, at some significant computational cost, by using some standard hyperparameters tuning methods (e.g., search grid and cross validation). The BART algorithm was included in this analysis because it is among the state of the art in machine learning algorithms, and is similar to gradient boosting machine algorithm that has shown good accuracy in predicting commercial buildings energy consumption (Touzani et al. 2018). Additionally, use of the default hyperparameters from Chipman et al. (2010) speeds its computation time with respect to the gradient boosting machine. The implementation of the BART algorithm used in this work is available within the bartmachine R package (Kapelner and Bleich 2013)

Bayesian additive regression trees Daily model (BART_Daily)

The BART_Daily model characterizes energy consumption using the BART algorithm and the following independent variables: day of the week (e.g., 1 for Monday and 7 for Sunday), the daily average outside air temperature, the standard deviation of the daily outside air temperature and a dummy variable that is equal to 1 if the considered day is a holiday and 0 if not.

Daily linear model (LM_Daily)

Using the same independent variables as BART_Daily model, this daily energy consumption model is fit using the ordinary least squares (OLS) regression algorithm. Note that unlike the BART_Daily model the days of the week are considered as a dummy variable, which means that for each day of the week a variable is created that is equal to 1 if the data point corresponds to the considered day and 0 if not. The mathematical form of the model is defined as follow

$$E_i = \alpha_0 + \alpha_1 \bar{T}_i + \alpha_2 sd(T_i) + \alpha_4 H + \sum_d \alpha_d D_d \qquad (2)$$

where $\bar{T}_i$ is the daily average outside air temperature, $sd(T_i)$ is the standard deviation of the daily outside air temperature, D_d are binary variable (dummy variable) corresponding to the day of the week and H is a dummy variable that is equal to 1 if the considered day is a holiday and 0 if not. This model was chosen as a representative example an OLS model that the uncertainty formulation in ASHRAE Guideline 14 was originally designed for.

2.2 Uncertainty estimation

2.2.1 Uncertainty estimation background

In addition to approximating the true regression function $f(\boldsymbol{x})$ (see equation 1) by an estimated function $\hat{f}(\boldsymbol{x})$ (i.e., baseline model), it is highly desirable for the M&V application to have a measure of confidence in the prediction provided by the model. The confidence is measured by quantifying the uncertainty surrounding the predictions. However, there are two types of

prediction uncertainty. The first is linked to the accuracy of the estimate $\hat{f}(\boldsymbol{x})$ in comparison to the true regression function $f(\boldsymbol{x})$, which corresponds to the distribution of the quantity $f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})$. In the regression literature this component of the uncertainty is known as the confidence interval (CI). The second type of the prediction uncertainty is linked to the accuracy of the estimate $\hat{f}(\boldsymbol{x})$ in comparison to the response y, which corresponds to the distribution of prediction error $y - \hat{f}(\boldsymbol{x})$. This type of uncertainty is called the prediction interval (PI). For M&V applications the PI is of more practical use than the confidence interval since it estimates the accuracy with which the baseline model predicts the observed response y, and not just the accuracy of the approximation of the true regression function $f(\boldsymbol{x})$. Note that the PI is wider than the CI since the prediction error can be defined as:

$$y_i - \hat{f}(\boldsymbol{x}_i) = [f(\boldsymbol{x}_i) - \hat{f}(\boldsymbol{x}_i)] + \varepsilon_i \qquad (3)$$

Assuming that the two error components in (3) are statistically independent, the variance of the prediction error σ_y^2 can be expressed as

$$\sigma_y^2(\boldsymbol{x}_i) = \sigma_{\hat{f}}^2(\boldsymbol{x}_i) + \sigma_\varepsilon^2 \qquad (4)$$

where $\sigma_{\hat{f}}^2$ is the variance of the model error around the true regression function $f(\boldsymbol{x})$.

Thus, given a confidence level CL of $100(1-\alpha)\%$ the PI of the response y_i corresponds to the interval that can be defined as $I_y^\alpha(\boldsymbol{x}_i) = \left[L_y^\alpha(\boldsymbol{x}_i), U_y^\alpha(\boldsymbol{x}_i)\right]$, where $L_y^\alpha(\boldsymbol{x}_i)$ is the lower bound and $U_y^\alpha(\boldsymbol{x}_i)$ the upper bound, which can be defined as

$$L_y^\alpha(\boldsymbol{x}_i) = \hat{f}(\boldsymbol{x}_i) - t_{1-\alpha/2,df}\sqrt{\sigma_y^2(\boldsymbol{x}_i)} \qquad (5)$$

$$U_y^\alpha(\boldsymbol{x}_i) = \hat{f}(\boldsymbol{x}_i) + t_{1-\alpha/2,df}\sqrt{\sigma_y^2(\boldsymbol{x}_i)} \qquad (6)$$

where $t_{1-\alpha/2,df}$ is the $1-\alpha/2$ quantile of a cumulative t-distribution function with df degrees of freedom. This statistical metric is also known as the t-score, or critical point of the t-distribution with df degrees of freedom. When the degrees of freedom exceeds 100, which roughly speaking corresponds to a number of pre period observations higher than 100, the $t_{1-\alpha/2,df}$ metric converges to the $z_{1-\alpha/2}$, which is the $1-\alpha/2$ quantile of a cumulative standard normal distribution (also known as z-score).

For example, using the assumption that underlie the ASHRAE Guideline 14 savings uncertainty quantification formula, and which is: the only input variable that is considered for the regression is the outside air temperature T and that $\hat{f}(T_i)$ is estimated by a linear regression model using the ordinary least squares method (OLS), the $\sigma_y^2(T_i)$ is estimated by

$$\hat{\sigma}_y^2(T_i) = s^2\left(1 + \frac{1}{n} + \frac{(T_i - \bar{T})^2}{\sum_{j=1}^{n}(T_j - \bar{T})^2}\right) \qquad (7)$$

where $\bar{T}$ is the sample mean and s^2 is an unbiased estimate of σ_ε^2 given by the mean squared error:

$$s^2 = MSE = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{f}(T_i))^2 \qquad (8)$$

2.2.2 Uncertainty estimation approaches

Two different approaches to quantify the uncertainty in savings estimates due to model error were evaluated in this work. The first is the M&V practitioners' standard defined in ASHRAE Guideline 14. The second is based on the cross validation method used by the statistics and machine learning communities to estimate the prediction accuracy of machine learning models.

ASHRAE Guideline 14 approach

ASHRAE Guideline 14 provides an equation to estimate the baseline model uncertainty of the savings estimates; that is, it provides a formulation to estimate the prediction interval surrounding the total savings over the post-installation period. This formulation was originally introduced by Reddy and Claridge (2000), and it is derived from the definition of the variance of the prediction error σ_y^2 for an OLS based baseline model with outside air temperature as independent variable (Equation 7). More specifically the ASHRAE Guideline 14 equation is an approximation of the aggregation of the variance of the prediction errors σ_y^2 (Equation 7) over *m* prediction points of the post period:

$$\Delta E_{save}^{ASHRAE} = 1.26\, t_{1-\alpha/2,df} \frac{\hat{E}_{post}}{m\bar{E}_{pre}} \sqrt{MSE(1+\frac{2}{n})m} \qquad (9)$$

where ΔE_{save}^{ASHRAE} is the uncertainty in the aggregated savings, *n* is the number of observations (data points) in the pre period, *m* is the number of observations in the post period, $\bar{E}_{pre}$ is the mean of the actual energy consumption in the pre period, $\hat{E}_{post}$ is the estimated energy consumption in the post period, 1.26 is an empirical factor of approximation provided by Reddy and Claridge (2000) in order to avoid the matrix algebra of the original equation of the aggregated uncertainties, MSE is the mean squared error of the baseline regression model defined as $MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(E_i^{pre} - \hat{E}_i^{pre}\right)^2$ with E_i^{pre} is the actual energy consumption is the pre-installation period, $\hat{E}_i^{pre}$ is the estimated energy consumption in the pre-installation period, $t_{1-\frac{\alpha}{2},n-k}$ is the t-statistic value with α the confidence level and df the degree of freedom. Note that typically within the M&V framework the ASHRAE formula (equation 9) is normalized by the savings estimate, this form of presenting the uncertainty is called fractional savings (ASHRAE 2014).

k-fold cross validation approach

In a more general case where no assumption is made on the type of baseline model used, one can estimate the prediction error σ_y^2, using k-fold cross validation (k-fold CV). The k-fold cross validation method consists of randomly splitting the training dataset (i.e., pre period) into *k* subsamples, called *folds*, of roughly equal size. In the first iteration, the baseline model is created using *k-1* folds as a training dataset and the held-out fold (prediction set) is used to calculate the first iteration estimation of the prediction error σ_y^2; this uncertainty is estimated using the mean squared error of the held-out fold (MSE_i). This procedure is repeated k times, and at each time a different fold is used as a test set. Figure 2 depicts an example of the k-folds CV approach, where k = 5. The final k-fold CV estimate of the MSE is the average of the MSE_i across each of the k iterations:

$$(\hat{\sigma}_y^2)^{CV} \approx \mathrm{MSE}^{CV} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \qquad (10)$$

where $MSE_i = \frac{1}{n_i}\sum_j^{n_i}(E_j^{pre} - \hat{E}_j^{pre})^2$ and n_i number of data points in the i-th held-out fold.

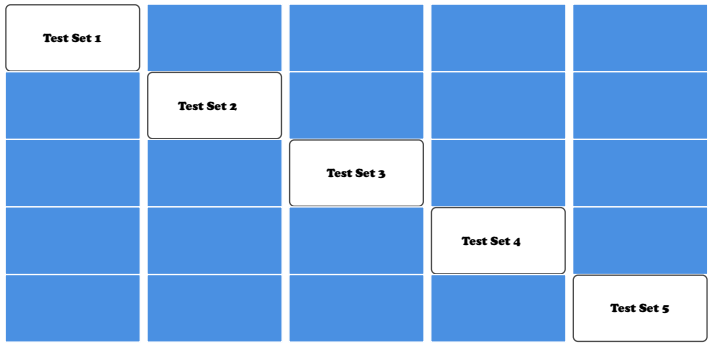


Figure 2. A k-fold cross validation, where k = 5. The blue boxes represent the training sets, and the white boxes represent the held-out sets used to estimate the MSE.

The advantage of using the k-folds CV to estimate the σ_y^2 versus the standard MSE calculation, which is the goodness of fit metric, is that the k-fold CV MSE version is computed using data points that were not included in the baseline model training process; as such, k-fold CV MSE should provide a better indication of the actual prediction uncertainty of the model and consequently a better estimate of the variance of the prediction error σ_y^2 (Equation 4). A modification of the traditional k-fold was applied, that defined the hold-out periods in a way that minimizes the effects of serial correlation at the borders of the training and test periods, and prevents inappropriate differences between the training and test periods (such as might occur if training on summer data).

Traditionally, in the time series forecasting literature, the out-of-sample evaluation method is used instead of the k-folds CV, where a block of data at the end of the time series is held out for testing. However, this can be problematic because the error can be completely misestimated if the statistical properties of the time series of the test set are different from the training data. To account for these issues, a modified version of the k-folds CV was used. This method consists in randomly selecting blocks of data points rather than randomly selecting unique observation (i.e., time step), when the k splitting is performed. More precisely, calendar weeks are considered as definition of a block. Thus, in the case where k=5, at each step of the 5-folds algorithm one fifth of available calendar weeks are selected as test sample and in addition in order to exclude the autocorrelation that may occur at the border (early hours of Mondays and late hours of Sundays) the Sundays preceding the test weeks and the Mondays following the test weeks are excluded from the training period. For example, in the case where 50 baseline weeks were available, at each iteration of the k-folds CV algorithm 10 weeks will be randomly chosen as held-out fold (prediction set) and the remain 40 weeks, not including the days that surround the held-out weeks, will be used to train the baseline model.

Although, there is no formal rule to choose the value of k, in practice k = 5 or k = 10 is used (James et al. 2013). As the value of k increases, the computational demands also increase. Therefore, to decrease the required computational time k =5 was applied in this work.

Following similar assumptions to ASHRAE Guideline 14, namely that the errors are independent and normally distributed, the uncertainty of the cumulative savings (i.e., error propagation definition, when the errors are independent and normally distributed) using the k-fold CV method can be estimated as:

$$\Delta E_{save}^{CV} = z_{1-\alpha/2}\sqrt{m\, MSE^{CV}} \qquad (11)$$

where m is the number of data points in the post-installation period.

One of the major differences between the k-fold CV method (Equation 11) and the ASHRAE formulation (Equation 9) is that the k-fold CV approach provides a non-deterministic estimate of the uncertainty, while the ASHRAE approach is deterministic. Thus, different trials will produce different estimates of MSE^{CV}. In this work, we are reporting only one trial estimate MSE^{CV}. Note that it is often advocated to repeat the k-fold cross validation and to average the results. However, as it is argued in (Vanwinckelen and Blockeel 2012) repeated cross validation does not necessarily provide much more accurate estimate of the model accuracy.

Adjustment for autocorrelation

When high resolution meter data is used (e.g., 15-min, hourly or daily) autocorrelated model errors arise. In other words, the time series of the errors is not random in time, and the information in each error observation is not totally separate from the information in other error observations. Consequently, the number of independent error observations is fewer than n (the number of observations in the pre period). The presence of the autocorrelation is usually induced by the omission of time dependent variables from the baseline model. These may be unknown, or not easily or cost-effectively measured. For example, occupancy is a variable that is known to have a significant impact on the energy use of a building and it is usually time dependent in commercial buildings, however it is very uncommon to have access to time-resolved measures of building occupancy levels.

The reduction in number of independent observations has implications on the uncertainty estimation such that the formulas 9 and 11 are no longer adequate. For positive autocorrelation, which is generally the case for commercial building energy use data, the MSE estimate of the errors variance will underestimate the true variance, and therefore underestimate the uncertainty in the savings estimate. ASHRAE's Guideline 14 introduced a version of ΔE_{save}^{ASHRAE} that corrects for autocorrelation by adjusting the number of observations in the pre-installation period. The true number of pre-installation period observations *n* is replaced by a quantity known as the effective number of observations *n'* (also called the number of independent observations). This correction is deterministic, based on the assumption that the autocorrelation in the time series is first-order autocorrelation, meaning that it is characterized by lag 1 or that autocorrelation is present only between consecutive values in the time series. The computation of the effective number of observations requires only the number of pre period observations and the lag 1 autocorrelation coefficient of the baseline model errors:

$$n' = n \cdot \frac{1-\rho}{1+\rho} \qquad (12)$$

The corrected version of the uncertainty in the aggregated savings, defined in ASHRAE Guideline 14, is expressed as

$$\Delta E_{save}^{ASHRAE} = 1.26\, t_{1-\alpha/2,df} \frac{\hat{E}_{post}}{m\bar{E}_{pre}} \sqrt{MSE \frac{n}{n'}(1+\frac{2}{n'})m} \qquad (13)$$

In order to account for autocorrelation of lag 1, the same deterministic method as use in the ASHRAE Guideline is applied to the k-fold cross validation approach. Thus, the corrected version of (11) is expressed as

$$\Delta E_{save}^{CV} = z_{1-\alpha/2} \sqrt{m \frac{n}{n'}\, MSE^{CV}} \qquad (14)$$

2.3 Method to evaluate uncertainty estimation approaches

The methodology that was developed to evaluate uncertainty estimation approaches is similar to the procedure that was used in Granderson and Price (2014), Granderson et al (2015) and Granderson et al (2016) to evaluate predictive accuracy of baseline models. This 4-step methodology is defined as follows:

1) Collect a dataset that comprises interval meter data and independent variable data, which is outside air temperature, for several hundred buildings. These buildings are "untreated" in terms of efficiency interventions. That is, they are not known to have implemented major efficiency measures.
2) The data for each building is divided into hypothetical training periods and prediction periods, and meter data from the prediction period is "hidden" from the model. For this study 6-month and 12-month training periods were considered.
3) For the set of baseline models and uncertainty methods described in Sections 2 and 3, train the baseline models using the training period data and generate predictions ($\hat{E}$) using the prediction period data; also, estimate the corresponding uncertainty (Δ) in the prediction at the 95% confidence level.
4) To evaluate the performance of each uncertainty method, compare the actual total energy consumption in the prediction period (E) to determine whether it falls within $\hat{E}$ +/- Δ. If it does, the uncertainty estimate is informative.

To evaluate the absolute accuracy of a given uncertainty quantification approach, we define the error uncertainty ratio (EUR), that characterizes whether the actual total energy consumption during the prediction period lies within the uncertainty range of the predicted consumption. That is, the EUR evaluates the quality of the prediction interval of the predicted total energy consumption of the post-installation period. If EUR is within the range [-1,1] then the estimated value of the total energy consumption is within the range of the uncertainty (i.e., prediction interval). If the EUR is outside the range [-1,1] then the uncertainty is underestimated. The EUR is defined as:

$$EUR = \frac{E_{total} - \hat{E}_{total}}{\Delta} \qquad (15)$$

where E_{total} is the actual value of the total energy consumption during the prediction period, $\hat{E}_{total}$ is the predicted value and Δ is the estimated uncertainty.

Analyzing the EUR metric over a large set of buildings provides good insight as to the accuracy of the uncertainty estimate. We complement this with an additional metric that quantifies the

percentage of buildings in the data set for which the actual total energy consumption is within the uncertainty range. In other words, *UICF* represents how often (the frequency) the observed total energy consumption falls within the uncertainty interval. This metric is referred to as uncertainty interval coverage factor (UICF).

$$UICF = 100\frac{1}{N}\sum_{i=1}^{N} I(E_{total}) \quad (16)$$

where

$$I(E_{total}) = \begin{cases} 1 & if\ E_{total}\ is\ within\ the\ uncertainty\ interval \\ 0 & else \end{cases}$$

In the equation N is the number of buildings in the test dataset. If $UICF$ is close to the nominal confidence level, for instance 95%, then the uncertainty method provides an informative uncertainty estimate. If *UICF* is much lower than the nominal confidence level, then the uncertainty method provides poorly informative estimates. If the *UICF* is equal to 1 it is important to analyze the EUR metric in order to check that the uncertainty estimation method is not overestimating the uncertainty.

2.4 Test datasets

To answer the research question posed in this work regarding the uncertainty estimation, we use a dataset that comprises interval meter data of buildings that are "untreated" in terms of efficiency interventions. That is, they are not known to have implemented major efficiency measures. The dataset is divided into two intervals, a hypothetical baseline (i.e. model training) period, and a hypothetical performance, or measure 'post' (i.e. prediction) period. Using this type of dataset allows us to perform an evaluation of the how well the uncertainty in 'post', or prediction period consumption, and therefore in the savings, is being estimated; this is because the actual metered consumption is known in the prediction period.

The baseline models and uncertainty estimation approaches were used with a dataset that comprises 15-minute metered whole-building electricity data gathered from 69 commercial buildings. The buildings are located in Northern and Central California (n = 54) and Washington, D.C. (n = 15). All buildings from the dataset had 24 months of electricity consumption and outside air temperature data; missing data was present in some of the time series but amounted to less than one percent of the total number of observations. A visual inspection showed no anomalous changes in energy use that would confound model fitness predictions.

To decrease the computational time of the analysis and to align with the hourly frequency of the outside air temperature data, the 15-minute meter data were aggregated to hourly intervals. In addition, since daily baseline models were used in this study, the data were also converted to daily interval format. The outdoor air temperature data were acquired using the ZIP code of each building and the closest weather station from the Weather Underground service (wunderground 2017).

For each building, the time series data were split into a hypothetical training period and prediction period. The prediction period was defined as the most recent 12 months of the available data. A 12-month prediction period (post-period) is generally the standard for whole-building M&V of energy savings. The models were trained using the 12 months and 6 months worth of data that immediately preceded the prediction period.

3. Results

3.1 Baseline model accuracy

First, we assessed the overall predictive accuracy of the models in order to verify that the considered baseline models provide reliable predictions of the post period. As in the assessment of uncertainty approaches, and as described in Granderson et al (2015) and Granderson et al (2016) the models were trained using either 12 or 6 months of data, and run to predict energy consumption for a 12 month period. The predictive accuracy was evaluated using two different statistical metrics, which are the coefficient of variation of the root mean squared error (CV(RMSE)), and the normalized mean bias error (NMBE). These two metrics provide complementary views of model performance for M&V applications, and they are also well adapted to assess relative model-to-model comparisons across the test dataset. A more extensive description of these metrics in addition to the summary table of all the following figures are provided in the appendix.

Figure 3 summarizes the CV(RMSE) results across the full population of buildings in the test dataset. In these box-and-whisker plots, the right end of each 'whisker' represents the CV(RMSE) for the 90th percentile in the population of test buildings, and the left end represents the 10th percentile. The right and left ends of each box represent the 75th and 25th percentiles, respectively, and the vertical line in each box marks the median, or 50th percentile. The most accurate results were obtained by the daily models, especially by the BART_Daily model that produced baseline models with CV(RMSE) smaller that 10% for more than 75% of test buildings. When the training period was reduced from 12 months to 6 months the performance decreased just slightly.

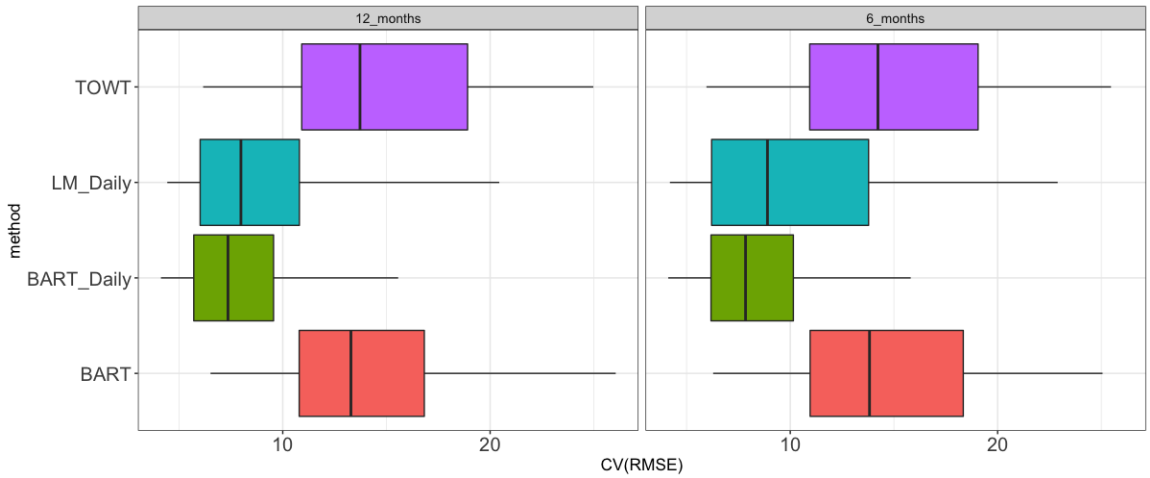


Figure 3. Distribution of CV(RMSE) metrics for each model, for 12-months prediction period, and 12-months and 6-months training period.

Figure 4 displays the distribution of NMBE across the test dataset, where the vertical red dashed line represents NMBE equal to 0. For the majority of cases there was a tendency of a bias toward over-predicting the energy consumption (NMBE negative). The results show that the differences in performance across the baseline models are modest. When the training period was reduced from 12 months to 6 months the performance of these models decreased.

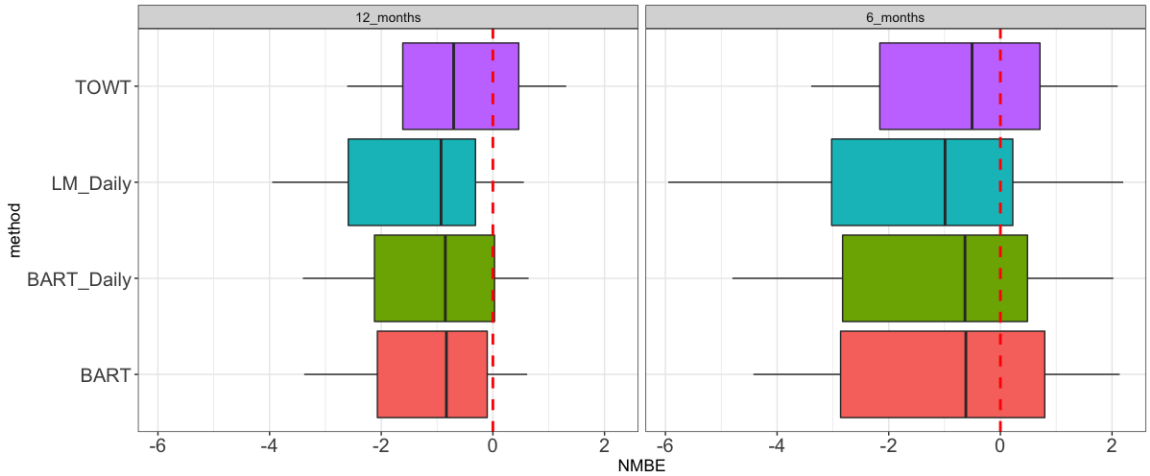


Figure 4. Distribution of NMBE metrics for each model, for 12-months prediction period, and 12-months and 6-months training period.

3.2 Uncertainty methods performance evaluation

To evaluate the absolute performance of each uncertainty quantification approach the error uncertainty ratio (EUR) across the full population of buildings in the test dataset is computed for each model and each uncertainty method. Thus, we analyze eight configurations of baseline model and uncertainty quantification method. The results of this analysis are shown in **Figure 5**. In the plots, CV stands for the k-fold CV uncertainty method. The vertical dashed red lines correspond to EUR equal to -1 and 1 and delimit the interval outside of which the estimated uncertainty is underestimated. The results show that the smallest deviation from the [-1;1] interval is obtained when LM_Daily baseline model is used. The second smallest deviation from the [-1;1] interval was obtained using another daily model (i.e., BART_Daily). The results obtained using hourly models (i.e., TOWT and BART) shows a significant deviation from the [-1;1] interval, which means that the estimated uncertainties are poorly informative.

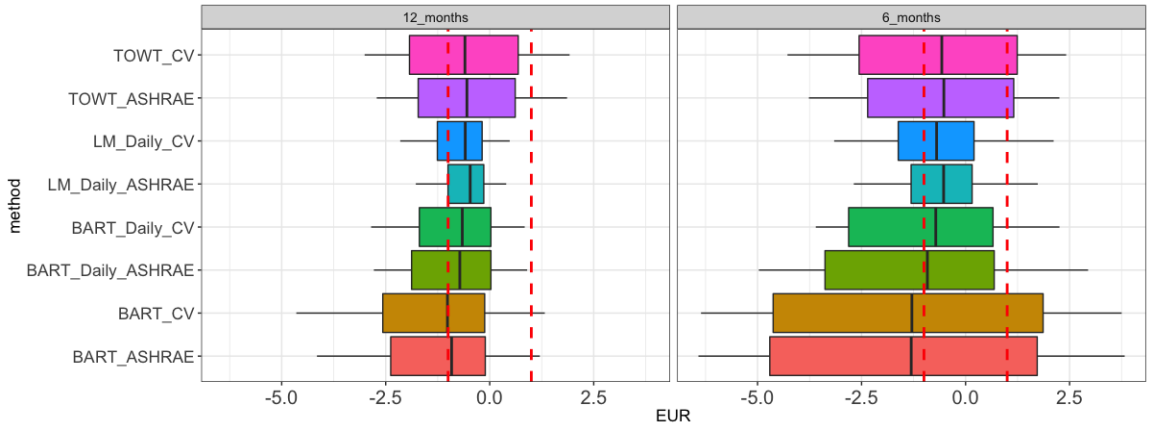
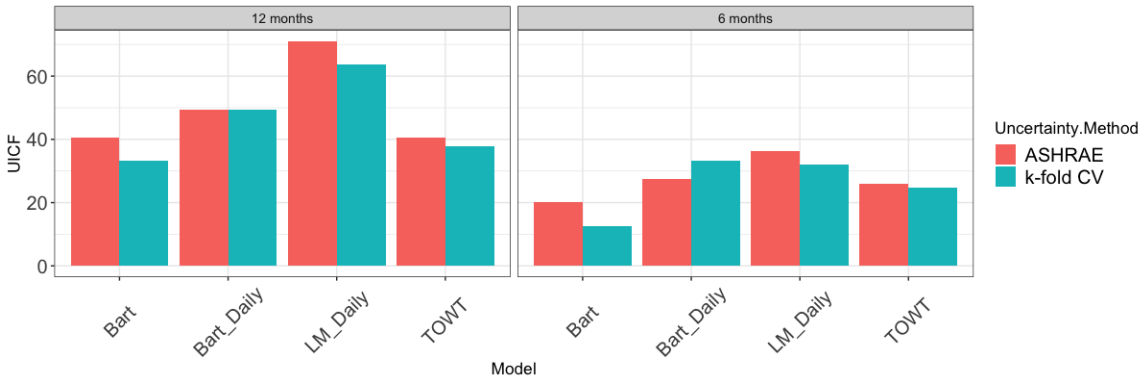


Figure 5. Distribution of EUR metrics for each combination of baseline model and uncertainty method, evaluated using the 12-months prediction period, and 12-months and 6-months training period.

Figure 6 summarizes the results of the computation of uncertainty interval coverage factor (UICF) for all combinations of baseline models and uncertainty methods and for both training

periods. The key finding is that the maximum UICP is obtained using the LM_Daily model, where the UICP using the ASHRAE uncertainty method is ~71% and using the k-folds CV method is ~64%. This means that for more than two thirds of buildings in the data set, the actual value of the annual energy consumption was within the estimated uncertainty interval. Meanwhile, for hourly models, the uncertainty methods provided an informative uncertainty interval for more than ~40% or fewer of the test buildings. However, these results are significantly below the expected 95%, which correspond to the 95% confidence level at which the uncertainties were computed.

Figure 6. UICF metrics (in percentage) for each combination of baseline model and uncertainty method, evaluated using the 12-months prediction period, and 12-months (left plot) and 6-months (right plot) training period.



4. Discussion

The results indicate that the approaches that were tested were not able to consistently provide informative estimates of the uncertainty in savings due to model error, when applied to higher frequency and non-linear models. The impact of higher frequency was more severe than that of using non-linear/non-parametric models, likely due to insufficient ability to correct for autocorrelation. The presence of the high degrees of autocorrelation in the model residuals (i.e., the differences between model predictions and actual data), violate the assumptions that underlie the uncertainty quantification approaches that were evaluated, and lead to underestimation of the uncertainty. The simple deterministic autocorrelations correction used in this work, and proposed in ASHRAE Guideline 14, is limited to correct autocorrelations of lag 1 and did not prove sufficient for either daily or hourly models, for the majority of buildings tested. In reality, the structure of the autocorrelation of the residuals is more complex than the lag 1 assumption, particularly for the residuals associated with hourly baseline models.

The ASHRAE approach used in combination with a daily linear model did stand out as providing the most informative uncertainty estimate. This was expected, as the ASHRAE Guideline 14 uncertainty formulation was designed for use with linear OLS approach. However, at its best, the

ASHRAE estimate was correct for 71% of the buildings analyzed, where the expected value was 95% (due to the 95% confidence level that was applied). In general, the k-folds cross validation approach underestimated the uncertainty with respect to the ASHRAE approach. When the duration of the training period was reduced from 12 months to six months, the number of correct estimates decreased significantly.

5. Conclusions

In this work we proposed and applied a methodology to evaluate uncertainty approaches that can be used to quantify the uncertainty of energy savings estimates, including the current industry standard approach defined in ASHRAE Guideline 14. While the ASHRAE approach was derived from first principles combined with simplifications that are valid for linear models with low serial correlation, one of the key contributions of this work was the development of a methodology to empirically test alternative uncertainty estimation approaches and their efficacy for different modeling techniques. The methods that were tested tend to underestimate the uncertainty. The tendency to underestimate the uncertainty is stronger for hourly models than for daily models, due to stronger autocorrelation in model residuals at the hourly time scale. The correction for autocorrelation that is proposed in ASHRAE Guideline 14, did not prove sufficient for either daily or hourly models, for the majority of buildings tested.

In conclusion, we note that savings uncertainty quantification is presented in the literature and in industry M&V references as a useful method to understand the impact of model error on the savings estimate. At the same time, higher frequency baseline models are being recognized as powerful approaches to quantify more resolved savings estimates, and to better characterize commercial building load profiles. Similarly, as solutions from the machine learning and data science fields are integrated into the building energy domain, more complex modeling approaches are increasingly being explored. With these trends as grounding context, recommendations from the results of this work, and prior research findings, is: where higher frequency meter data is available (daily or hourly, as opposed to monthly), it can be leveraged to improve load and energy use predictions. Metrics such as R^2, CV(RMSE), MSE, and NMBE are sound measures of model error, and therefore appropriate to assess model suitability for conducting savings measurement and verification. The most effective methods of quantifying these metrics are those that use cross validation, as opposed to simple goodness of fit, and are therefore more robust in estimating prediction accuracy.

In order to improve the accuracy of the post-installation period savings estimates and the corresponding uncertainty quantification, it is important that the M&V engineering and research community develop better methods to address the baseline models' errors autocorrelation. A potential solution, is to use autoregressive models to remove the autocorrelation, however, the existing autoregressive models are not directly applicable to the M&V problem of producing post savings estimates because these methods have been mostly developed for time series forecasting, which involves the usage of observations of previous timesteps to predict the value of the next time step. Thus, since there is generally a significant delay between the end of the baseline period and the beginning of the post-installation period, which correspond to the period of time where the measure is implemented in the building, it is impossible to use the previous timesteps to predict the energy use in the post period. This challenging research problem need to be resolved in order to use the potential of the autoregressive models to address the autocorrelation in the baseline models' errors.

Acknowledgment

This work described in this report was funded by the Pacific Gas and Electric Company. Lawrence Berkeley National Laboratory's contributions were also supported by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This project was developed as part of Pacific Gas and Electric Company's Emerging Technology – Technology Development Support program under internal project number: ET12PGE5312.

References

Afram, A. and Janabi-Sharifi, F., 2014. Theory and applications of HVAC control systems–A review of model predictive control (MPC). Building and Environment, 72, pp.343-355.

Ahmad, M.W., Mourshed, M. and Rezgui, Y., 2017. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings, 147, pp.77-89.

Araya, D.B., Grolinger, K., ElYamany, H.F., Capretz, M.A. and Bitsuamlak, G., 2017. An ensemble learning framework for anomaly detection in building energy consumption. Energy and Buildings, 144, pp.191-206.

ASHRAE Guideline 14 (2014). ASHRAE Guideline 14-2014 for Measurement of Energy and Demand Savings. American Society of Heating, Refrigeration and Air Conditioning Engineers, Atlanta, Georgia.

Chipman, H.A., George, E.I. and McCulloch, R.E., 1998. Bayesian CART model search. Journal of the American Statistical Association, 93(443), pp.935-948.

Chipman, H.A., George, E.I. and McCulloch, R.E., 2010. BART: Bayesian additive regression trees. The Annals of Applied Statistics, 4(1), pp.266-298.

Consortium for Energy Efficiency (CEE). State of the efficiency program industry: Budgets, expenditures, and impacts. March 2017, Consortium for Energy Efficiency.

Coulston, J.W., Blinn, C.E., Thomas, V.A. and Wynne, R.H., 2016. Approximating prediction uncertainty for random forest regression models. Photogrammetric Engineering & Remote Sensing, 82(3), pp.189-197.

Efron, B. and Tibshirani, R.J., 1994. An introduction to the bootstrap. CRC press.

Energy Information Administration (EIA). Commercial Buildings Energy Consumption Survey (CBECS): 2012 CBECS Survey Data. Energy Information Administration, May 2016. Available from: http://www.eia.gov/consumption/commercial/reports/2012/energyusage/index.cfm; accessed November 14, 2017.

EVO, 2016 Efficiency Valuation Organization (EVO) International Performance Measurement and Verification Protocol: Concepts and Options for Determining Energy and Water Savings, vol. 1, EVO (2016) 10000–1:2016

Fels, M.F., 1986. PRISM: an introduction. Energy and Buildings, 9(1), pp.5-18.

Granderson, J. and Price, P.N., 2014. Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models. Energy, 66, pp.981-990.

Granderson, J., P. N. Price, D. Jump, N. Addy, and M. D. Sohn. 2015. "Automated measurement and verification: Performance of public domain whole-building electric baseline models." Applied Energy, 144, pp.106–113.

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

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

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

Heo, Y. and Zavala, V.M., 2012. Gaussian process modeling for measurement and verification of building energy savings. Energy and Buildings, 53, pp.7-18.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Kapelner, A. and Bleich, J., 2013. bartmachine: Machine learning with Bayesian additive regression trees. arXiv preprint arXiv:1312.2171.

Kissock, J.K., Haberl, J.S. and Claridge, D.E., 2002. Development of a Toolkit for Calculating Linear, Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models, ASHRAE Research Project 1050-RP, Final Report. Energy Systems Laboratory, Texas A&M University.

Koran, W. E. 2017. Uncertainty Approaches and Analyses for Regression Models and ECAM, Bonneville Power Administration. Available from: https://www.bpa.gov/EE/Utility/research-archive/Documents/Evaluation/UncertaintyMethodsComparisonsFinal.pdf; accessed on December 12, 2017.

Koran, B., Boyer, E., Khawaja, S., Rushton, J., and Stewart, J, 2017. *A Comparison of Approaches to Estimating the Time-Aggregated Uncertainty of Savings Estimated from Meter Data*. IEPEC Proceedings, 2017.

Mamidi, S., Chang, Y.H. and Maheswaran, R., 2012, June. Improving building energy efficiency with a network of sensing, learning and prediction agents. In Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1 (pp. 45-52).

Mathieu, J. L., P. N. Price, S. Kiliccote, and M. A. Piette. 2011. "Quantifying changes in building electricity use, with application to demand response." IEEE Transactions on Smart Grid, 2(3), pp. 507–518.

Najafi, M., Auslander, D.M., Bartlett, P.L., Haves, P. and Sohn, M.D., 2012. Application of machine learning in the fault diagnostics of air handling units. Applied energy, 96, pp. 347-358.

Reddy, T.A. and Claridge, D.E., 2000. Uncertainty of "measured" energy savings from statistical baseline models. HVAC&R Research, 6(1), pp.3-20.

Stuart, E., Larsen, P.H., Carvallo, J.P., Goldman, C, and Gilligan, D. U.S. Energy service company (ESCO) industry: Recent market trends. October 2016, Lawrence Berkeley National Laboratory, LBNL Report #1006343.

Subbarao, K., Lei, Y. and Reddy, T.A., 2011. The Nearest Neighborhood Method to Improve Uncertainty Estimates in Statistical Building Energy Models. ASHRAE Transactions, 117(2).

Touzani, S., Granderson, J. and Fernandes, S., 2018. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, pp.1533-1543.

Touzani, S., and Granderson, J., 2017. R Package for Measurement and Verification 2.0 for Commercial Buildings (RMV2.0) v1.0. LBNL 2018-001. Available from: https://github.com/LBNL-ETA/RMV2.0

Vanwinckelen, G. and Blockeel, H., 2012, May. On estimating model accuracy with repeated cross-validation. In *Proceedings of the 21st Belgian-Dutch Conference on Machine Learning* (pp. 39-44).

Walter, T., Price, P.N. and Sohn, M.D., 2014. Uncertainty estimation improves energy measurement and verification procedures. Applied Energy, 130, pp.230-236.

wunderground. The Weather Channel LLC. 2017. Weather Underground. Available at: http://wunderground.com

Appendix

Statistical Metrics to Assess Model Accuracy

To evaluate the effectiveness of a baseline model, several statistical metrics can be used, and these different metrics provide different insights into aspects of accuracy measurement. Relying on just one metric is usually not sufficient to fully understand the weakness and strengths of a specific baseline model. The two metrics that are used in this work are the normalized mean bias error (NMBE); and the coefficient of variation of the root mean squared error (CV(RMSE)).

The NMBE is the mean of the error in the predictions divided by the mean of the actual energy use. In other words, it gives a sense of the total difference between model predicted energy consumption, and actual metered energy use, with intuitive implications for the accuracy of avoided energy use calculations. If the value of NMBE is positive, it means that the prediction of the total energy used during the entire prediction period is lower than the measured value. A negative NMBE means that the prediction is higher. The NMBE is defined in the following equation, where $\bar{y}$ is the average of y_i.

$$NMBE = \frac{\frac{1}{n}\sum_i^n (y_i - \hat{y}_i)}{\bar{y}} \times 100$$

The value of NMBE is independent of the timescale for which it is evaluated, which means that the value of the metric will be the same if the timescale is 15-minute, hourly, or daily.

The CV(RMSE) is the root mean square error normalized by the mean of the measured values, which provides a quantification of the typical size of the error relative to the mean of the observations. This metric also gives an indication of the model's ability to predict the overall energy use shape that is reflected in the data. CV(RMSE) is also familiar to practitioners, and is prominent in resources such as ASHRAE Guideline 14. The CV(RMSE) is defined by the equations below, where y_i is the actual metered value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the average of the y_i, and n is the total number of data points.

$$CV(RMSE) = \frac{\sqrt{\frac{1}{n}\sum_i^n (y_i - \hat{y}_i)^2}}{\bar{y}} \times 100$$

In contrast to the NMBE, CV(RMSE) quantify the predictive accuracy at the timescale of the data and prediction; in other words, if the predictions and measured data apply to hourly timescales, then this metric summarizes the accuracy in hourly predictions.

Results Summary Tables

Model	P10	P25	P50	P75	P90
BART	6.3	11	14	18	25
BART_Daily	4.1	6.2	7.8	10	16
LM_Daily	4.2	6.2	8.9	14	23
TOWT	6	11	14	19	25

Table A-1. Percentiles of CV(RMSE) metrics for each baseline model, evaluated using the 12-months prediction period, and 6-months training period.

Model	P10	P25	P50	P75	P90
BART	6.5	11	13	17	26
BART_Daily	4.1	5.7	7.4	9.6	16
LM_Daily	4.4	6	8	11	20
TOWT	6.2	11	14	19	25

Table A-2. Percentiles of CV(RMSE) metrics for each baseline model, evaluated using the 12-months prediction period, and 12-months training period.

Model	P10	P25	P50	P75	P90
BART	-4.4	-2.9	-0.62	0.8	2.1
BART_Daily	-4.8	-2.8	-0.63	0.49	2
LM_Daily	-5.9	-3	-0.99	0.22	2.2
TOWT	-3.4	-2.2	-0.51	0.71	2.1

Table A-3. Percentiles of NMBE metrics for each baseline model, evaluated using the 12-months prediction period, and 6-months training period.

Model	P10	P25	P50	P75	P90
BART	-3.4	-2.1	-0.83	-0.1	0.61
BART_Daily	-3.4	-2.1	-0.85	0.03	0.64
LM_Daily	-4	-2.6	-0.93	-0.31	0.55
TOWT	-2.6	-1.6	-0.7	0.46	1.3

Table A-4. Percentiles of NMBE metrics for each baseline model, evaluated using the 12-months prediction period, and 12-months training period.

Method	P10	P25	P50	P75	P90
BART_ASHRAE	-6.4	-4.7	-1.3	1.7	3.8
BART_CV	-6.4	-4.6	-1.3	1.9	3.8
BART_Daily_ASHRAE	-5	-3.4	-0.92	0.69	2.9
BART_Daily_CV	-3.6	-2.8	-0.72	0.66	2.3
LM_Daily_ASHRAE	-2.7	-1.3	-0.53	0.16	1.7
LM_Daily_CV	-3.2	-1.6	-0.7	0.2	2.1
TOWT_ASHRAE	-3.8	-2.4	-0.52	1.2	2.3
TOWT_CV	-4.3	-2.6	-0.57	1.2	2.4

Table A-5. Percentiles of EUR metrics for each combination of baseline model and uncertainty method, evaluated using the 12-months prediction period, and 6-months training period.

Method	P10	P25	P50	P75	P90
BART_ASHRAE	-4.2	-2.4	-0.92	-0.11	1.2
BART_CV	-4.6	-2.6	-1	-0.12	1.3
BART_Daily_ASHRAE	-2.8	-1.9	-0.72	0.029	0.9
BART_Daily_CV	-2.9	-1.7	-0.66	0.028	0.84
LM_Daily_ASHRAE	-1.8	-1	-0.47	-0.14	0.4
LM_Daily_CV	-2.2	-1.3	-0.59	-0.18	0.48
TOWT_ASHRAE	-2.7	-1.7	-0.55	0.61	1.9
TOWT_CV	-3	-1.9	-0.6	0.69	1.9

Table A-6. Percentiles of EUR metrics for each combination of baseline model and uncertainty method, evaluated using the 12-months prediction period, and 12-months training period.

Model	ASHRAE		k-folds CV	
	12 months	6 months	12 months	6 months
BART	40.6	20,3	33.3	21.7
BART_Daily	49.3	27.5	49.3	33.3
TOWT	40.6	26.1	37.7	24.6
LM_Daily	71	36.2	63.7	31.9

Table A-7. UICF metrics (in percentage) for each combination of baseline model and uncertainty method, evaluated using the 12-months prediction period, and 12-months (left plot) and 6-months (right plot) training period.