Energy and Buildings 128 (2016) 1–11

Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

Predicting future monthly residential energy consumption using building characteristics and climate data: A statistical learning approach

Kristopher T. Williams*, Juan D. Gomez

Texas Sustainable Energy Research Institute, The University of Texas at San Antonio, San Antonio, TX, USA

ARTICLE INFO

Article history:
Received 10 May 2016
Received in revised form 21 June 2016
Accepted 24 June 2016
Available online 29 June 2016

Keywords:
Forecasting energy
Multivariate adaptive regression splines
Regression trees
Model validation

ABSTRACT

In this paper a large-scale study is presented that applies statistical learning methods to predict future monthly energy consumption for single-family detached homes using building attributes and monthly climate data. Building data is collected from over 426,305 homes in Bexar County, TX with four years of monthly energy consumption (natural gas and electricity). The goal of this study is to establish reliable models for forecasting residential energy consumption, understand the predictive value of building attributes, identify differences in predictability between households, and measure the robustness in model performance given uncertainty in climate forecasts. Assuming accurate climate forecasts, results show future monthly energy consumption can reasonably be predicted for out-of-sample households, with 74% accuracy at the household level and over 90% accuracy for predicting aggregate monthly energy usage. However, model performance is significantly different between households with distinct fuel types. Using historical climate forecast, results also demonstrate that model predictability significantly decays at both the household and aggregate level, but is robust at the household level when measured by the median home. Model selection and variable importance plots illustrate several building characteristics significantly contribute to predicting monthly energy consumption while most provide marginal predictive value.

Published by Elsevier B.V.

1. Introduction

Developing reliable models for predicting building energy is a challenging task. A recent study has highlighted the disparity between model forecasts and measured energy and the need to develop better predictive models [1]. The residential energy sector is one of the most difficult to predict on a household basis due to the large variability in energy usage between households [2]. Residential energy accounts for 21% of all energy consumption in the United States [3]. More specifically, single-family detached homes make-up more than 64% of all households and represent about 75% of all residential energy usage [3]. Given these trends, accurate forecasting of energy consumption for single-family detached homes is crucial for optimizing allocation of energy resources, protecting future energy supply/demand and promoting efficiency and conservation efforts. Another motivation for this paper is the lack of large scale validation studies within the energy research community [1]. Analyses on energy consumption are mostly from an inferential perspective, and studies that validate models normally do so using simulated data. Without validation results from real (e.g., not simulated) data the predictive value of many techniques may not truly be known

Before considering modeling techniques it is necessary to understand what factors contribute to residential energy consumption. Well-known factors significantly impacting residential energy include; climate, building characteristics, demographics and household behavior [4]. An extensive amount of literature exist examining the role these factors have on residential energy, such as the research by [5–9]. Building characteristics and climate are mostly publicly available while household behavior and demographic data are usually proprietary. Therefore, developing robust statistical models to predict energy consumption using readily available public data, provides a value added services to many institutions that may not have access to the full range of granular household information.

The goal of this study is to reliably predict future monthly household energy consumption for single-family detached homes as a

* Corresponding author.
E-mail addresses: kristopher.williams@utsa.edu (K.T. Williams), juan.gomez@utsa.edu (J.D. Gomez).

http://dx.doi.org/10.1016/j.enbuild.2016.06.076
0378-7788/Published by Elsevier B.V.

function of building and climate attributes, and provide a framework for understanding the predictive value of each attribute. Some of the research questions are: (1) Can monthly household energy consumption be reasonably predicted using building characteristics and climate? (2) What attributes are the most important for predicting energy consumption? (3) What types of homes are more difficult to predict? (4) What modeling techniques are most effective at predicting monthly household energy usage? (5) How does model performance respond to uncertainty in climate forecasts? To answer these questions three different statistical learning techniques are investigated: linear regression (LR), regression trees (RT), and multivariate adaptive regression splines (MARS). Linear regression is relatively simple to implement and provides easy to interpret estimates, and regression trees and MARS are capable of modeling more complex relationships among predictor variables with less interpretability. The method sections explain the details of each modeling technique. The data set used in analysis consisted of data combined from a variety of sources. Monthly natural gas and electricity usage is merged with publicly available building and climate information. Extensive validation results with model performance metrics are reported at the household and aggregate levels. To demonstrate the robustness of each modeling technique under uncertainty in climate predictions, results are also provided using historical climate patterns. Model selection procedures along with variable importance plots are utilized to identify important predictor features.

2. Related work

Predicting residential energy consumption can be accomplished in various contexts over different time frames (e.g., hourly, monthly, yearly). Swan and Ugursal [4] point out the main distinction between modeling methods, is the level of detail of input data. Input data can include building characteristics (size, vintage, fuel type, construction type, etc), historical energy consumption, appliance and electronic energy usage, demographics (income, race, gender, education, number of occupants, etc.), and climate data (temperature, humidity, etc.). Obtaining household energy consumption data at finer granularity is valuable for prediction, but rarely used due to the high costs associated with collecting such information [4]. However, with the increasing deployment of smart metering technology, more detailed household energy data is becoming available. Some authors have already applied modeling techniques to predict household energy consumption using minute and hourly household energy data [10,11], but these analyses use a relatively small sample size (on the order of 10s of households or less). Nonetheless, these studies provide insights into advanced modeling techniques for future analysis.

Recent empirical studies have tackled the challenge of predicting energy consumption for both monthly and annual usage using large data sets. Hosgor and Fischbeck [12], predict monthly residential energy using building, demographic, and climate data. Linear regression models are built using 3 years of monthly energy data (2009–2011) from over 10,000 single-family detached homes in Florida. The study reports that many predictor variables are significant; however, no validation results are reported to support their claims. For instance, the authors claim that political party affiliation has significant influence on the monthly energy consumption of households. This claim is based on the statistically significant p-values in a linear regression model associated with the political party affiliation predictor. Even though political party affiliation may be relevant to modeling energy consumption, relying only on p-values without further analysis can give misleading conclusions. In the paper by Kolter and Ferreira [13], annual building energy usage (commercial and residential) is predicted using building characteristics from 6500 buildings with several years of monthly energy consumption data. Models are validated on test data and the nonparametric technique called Gaussian regression is shown to be more effective compared to linear regression. Results from their study indicate that several building characteristics are significant predictors, such as; building appraisal value, size of living area, fuel type, and building style, while most building characteristics provide little predictive value. In other related studies, Dong et al. [14] show that support vector machines (SVMs) can accurately predict monthly building energy consumption for four commercial buildings using climate data, and Catalina et al. [15] demonstrate the validity of using neural networks over linear regression to predict monthly heating demand for residential buildings using simulated building and climate data. In this paper several contributions are presented that build on the previous mentioned literature. The remaining sections of the paper go as follows: an overview of each statistical learning method is introduced, data collection and data processing is described, and then model selection and validation results are provided along with a discussion.

3. Statistical learning methods

One goal of this study is to understand the effectiveness of different statistical learning techniques to predict residential energy consumption. While there are numerous statistical modeling approaches, three well-known methods are applied: linear regression, regression trees, and multivariate adaptive regression splines (MARS). Since linear regression is the simplest technique and cannot accommodate more complex nonlinear relationships, it is considered as a baseline model to compare all others. Given a training set $\{X, y\} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where X is a set of n, p-dimensional feature vectors in R^p and y is a one-dimensional response vector, the goal in supervised learning is to create a function f: $X \rightarrow y$ that most accurately maps the input feature vectors, X, to the output response, y.

3.1. Linear regression (LR)

One of the most popular regression techniques is linear regression. The linear regression model is given by

$$y = f(X) = X\beta + \epsilon \tag{1}$$

where β is the $p+1$ dimensional vector of coefficients and ϵ represents the set of n error terms. To estimate the coefficients, β, ordinary least squares is the most common method; although, maximum likelihood is regularly used. For maximum likelihood, different likelihood functions could be constructed based on assumption about the underlying distribution of the error terms, which may results in different maximum likelihood estimates of β. In this study, all linear regression estimates are obtained using ordinary least squares and inferential statistics reported are based on the assumption that $\epsilon_i \sim N(0, \sigma^2)$ iid, for $i = 1, \ldots, n$. Model selection is performed using backward elimination where the "best" model is selected using average 10-fold cross-validated R squared (R^2).

3.2. Regression trees (RT)

Classification and regression trees (CART) is a popular supervised learning technique that has been extensively applied in many domains including modeling building energy demand [10]. CART is a flexible technique that can model complex nonlinear relationships and high order interactions among predictor features. Originally developed by Breiman et al. [16], the CART model is a recursive partition method that finds the "best" disjoint regions of the data space to make predictions given an input, x. For regression trees, the model for x is written as:

$$f(x) = \frac{1}{n_c}\sum_{i \in c} y_i \mathbf{1}(x \in C) \tag{2}$$

Here the predicted value of x is simply the sample mean response in partition C, where C is the partition or terminal node x belongs to, n_c is the size of the node, and c is the set of indexes in C. The construction of the terminal nodes ($C_1, \ldots, C_m$) is described as follows. Given all points in X, a search over all binary splits of each predictor feature is done where the split which minimizes the error function (i.e., mean squared error) is selected as the "best" split. Once the "best" split is found, the data is divided into two nodes or regions. This process is repeated recursively for each successive region. The algorithm stops when a tree is fully grown based on a stopping criteria, like maximum tree size. Typically, the fully grown tree over-fits the data, so pruning methods are carried out using cross-validation. In this study, the fully grown tree is trimmed to the tree size with the optimal complexity parameter value, α. The optimal α value is obtained by selecting the α that corresponds with the 10-fold cross-validated error within one standard deviation of the minimum error. For further details of regression trees and the CART algorithm refer to [16]. Implementation of the algorithm is carried out using the "rpart" package from R open source software [17].

3.3. Multivariate adaptive regression splines (MARS)

MARS is a nonparametric regression technique that can be viewed as generalization of the additive model [18]. MARS has the ability to model nonlinear relationships between predictors and the response, along with interactions between predictors. First proposed by Friedman [19], MARS has been applied to many learning tasks such as predicting building energy performance [20]. The MARS model formula is expressed as:

$$y = f_M(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \tag{3}$$

where the response is modeled as a weighted sum of basis functions plus the intercept term. Here β_m is the mth weight, h_m is the mth basis function, and β_0 represents the intercept. To form the basis functions MARS uses the truncated basis functions

$$(x-t)_+ = \begin{cases} x-t & x > t \\ 0 & otherwise \end{cases}$$

where t represents the knot location of predictor x. The truncated basis functions divide the data into regions for approximating the true underlying function, f. To construct the final model, MARS performs two basic steps. The first step is a forward selection process that selects all possible basis functions and corresponding knots for all predictors and interaction terms specified by the user. The second step, is a backward elimination procedure that prunes the model by removing terms that increase the error function (mean squared error) the least. The pruning process stops when the generalized cross-validated (GCV) error function is minimized. GCV takes into account both the model error and number of terms in the model and is expressed as the ratio of cross-validated mean squared error divided by a penalty term. GCV is written as:

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_M(x_i)\right)^2}{\left(1 - \frac{C(M)}{n}\right)^2} \tag{4}$$

where $\hat{f}_M$ is the predicted response, M is the number of terms in the model and C is the complexity function expressed as, $C(M) = M(\frac{d}{2}+1)+1$. The parameter d, acts as a smoothing parameter, where larger values correspond to fewer basis functions. The goal of the pruning procedure is to select the model that has the least complexity (i.e., smallest number of model terms) and smallest error. All MARS model selection in this paper is done using 10-fold cross-validation. For a more thorough introduction to MARS, refer to [21]. The algorithm is implemented using the "earth" package from R open source software [22].

4. Data description and processing

4.1. Monthly energy consumption and fuel type

Monthly site energy consumption, both electricity and natural gas, is collected from the local electric utility from January 2010 to December 2013, for 426,305 single-family detached homes in Bexar County, Texas. Total monthly energy for each household represents monthly electricity (kBtu) plus gas (kBtu) consumption. To standardize energy consumption by month, the total monthly energy per household is divided by the number of days in a given month, which represents the average daily energy consumption (kBtu/day). Fig. 1 shows the box plot of average daily household energy consumption by month across all homes with non-zero monthly energy consumption for 2010–2013 (note: for visual purposes homes with abnormally high average daily energy consumption are removed). The final response variable is the log transformed average daily energy consumption. Taking the log controls for skewness in the distribution. After log scaling, the average daily energy consumption is approximately normally distributed, which is seen in Fig. 2.

An important feature constructed for analysis is the fuel type. Fuel type describes what type of energy (electricity or natural gas) is used to power the home. Fuel type categories are developed based on a detailed analysis of the electricity and natural gas consumption patterns. Five distinct profile patterns are identified to classify homes based on fuel type. The fuel type categories are described as follows: all-electric homes (fuel type 1), homes with access to gas but not for space heating (fuel type 2), homes with access to gas for space heating, water heating, and cooking (fuel type 3), homes with access to gas but not for water heating (fuel type 4), and homes with access to gas only for space heating (fuel type 5). From Fig. 3, the box plot of average yearly energy usage (kBtu/year) by fuel type, shows a clear distinction between all-electric and dual-fuel homes (except for fuel type 4) in terms of site energy consumption, with dual-fuel homes consuming about 27% more energy than all-electric homes. This contrast in energy consumption by fuel type is due to the fact that newer homes tend to be all-electric. From the data in this study, about 67% of single-family detached homes built in 1990 or after are all-electric, whereas, only 18.3% of homes built prior to 1990 are all-electric. Since newer homes are typically more efficient, this implies that all-electric homes are expected to consumer less site energy per square foot than dual-fuel homes across all vintages.

4.2. Building characteristics and climate data

Building characteristics are obtained from the local county property appraisal districts office. Table 1 summarizes the extracted building and land information used in this analysis. For climate data, hourly temperature and humidity measurements are collected from Weather Underground for readings at San Antonio International Airport. Climate data measured from this weather station is selected due to its historical rich database (over 30 years of hourly weather readings) and being centrally located within Bexar County, TX where all homes in this study reside. From the hourly readings, two climate features are created,

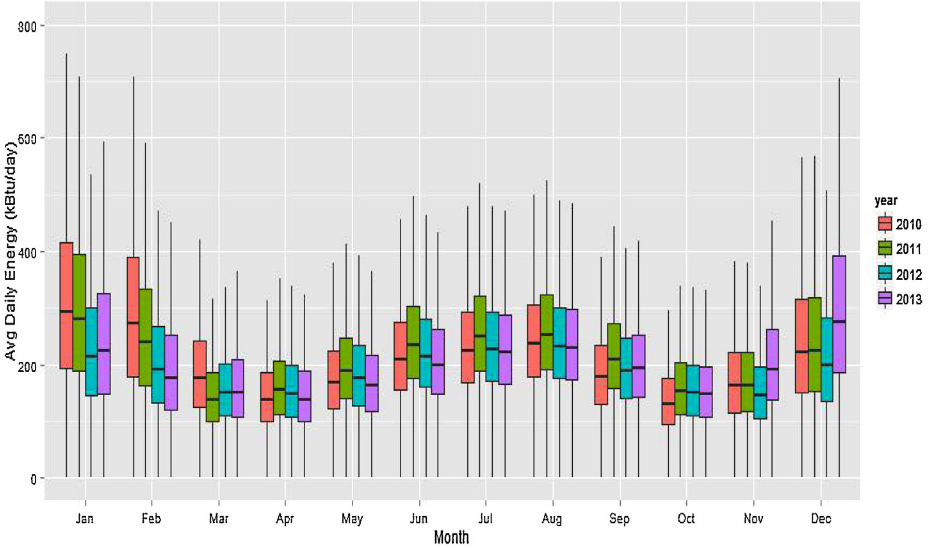

Fig. 1. Box plot of average daily energy consumption (kBtu/day) by month across all homes with non-zero energy consumption, 2010–2013.

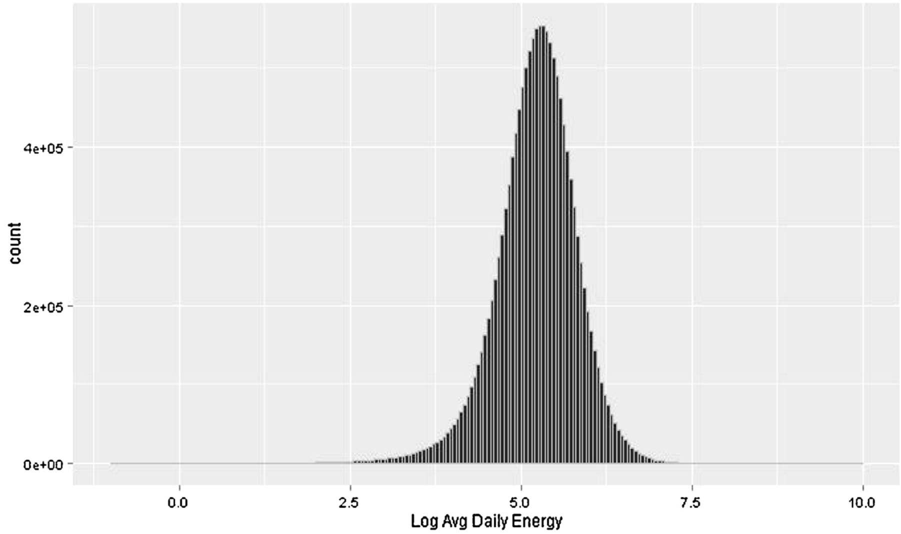

Fig. 2. Histogram of log average daily energy consumption (kBtu/day) across all homes with non-zero energy consumption for all months, 2010–2013.

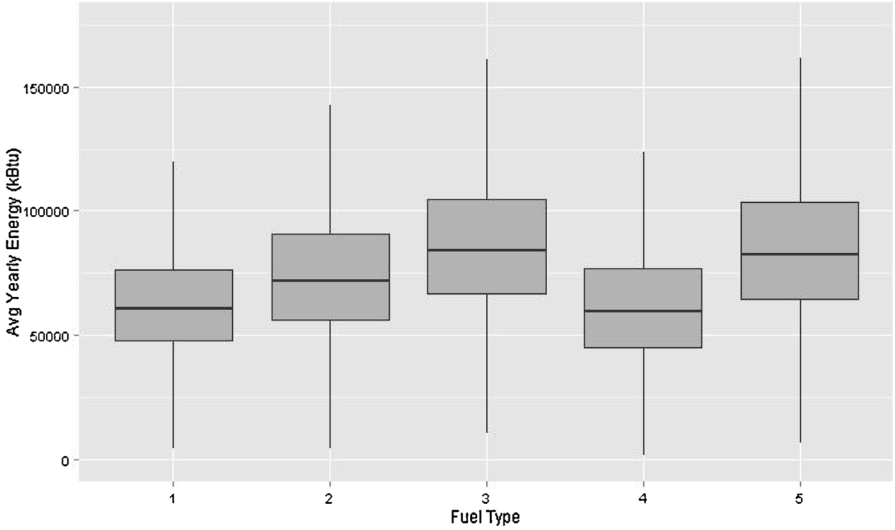

Fig. 3. Box plot of average yearly energy consumption (kBtu/year) by fuel type.

Table 1
Summary of building characteristics.

Description	Abbreviation
Size of Living Area (square footage -sf)	Size
Year Built	Vintage
Number of Rooms	No_Rooms
Number of Bedrooms	No_Bedrooms
Number of Bathrooms	No_Bathrooms
Number of Stories	No_Stories
Pool Ownership (yes or no)	Pool
Spa Ownership (yes or no)	Spa
Presence of Fireplace (yes or no)	Fireplace
Lot Size (square footage -sf)	Lot_Size
Construction Style (3 categories)	Constr_Style
Foundation Type (2 categories)	Foundation_Type
Exterior Wall Type (8 categories)	Ext_Wall
Roof Covering Type (6 categories)	Roof_Cov
Detached Living Area (yes or no)	Dtch_LA
Attached Porch (yes or no)	Att_Porch
Attached Garage (yes or no)	Att_Garage

average monthly temperature (F) and average monthly humidity (%). Average monthly temperature and average monthly humidity are defined as the sample mean of the average daily temperature and average daily humidity respectively. Fig. 4, shows the plot of the average monthly temperature and average monthly humidity for each year, 1983–2013.

4.3. Data processing

Before analysis, some homes are removed due to lack of data or inconsistencies with "normal" energy consumption patterns. Only homes built before 2010 that have electricity usage for all four years are considered, which represents 297,388 homes. Within this group, abnormally efficient/inefficient homes are removed. For this criteria, households with an average site energy use intensity (EUI = kBtu/sf/year) less than 10 kBtu/sf/year and greater than 140 kBtu/sf/year are removed. This represents the bottom 0.5% and top 99.5% of all households, relative to EUI. A plot of the average EUI across all households with non-zero monthly energy consumption is shown in Fig. 5. After selecting homes with the above criteria some building attributes have categories with low counts. The following rule is used to remove homes with a low attribute category: if an attribute category represents less than 1% of the population of homes then those homes within that attribute category are removed. The final set of homes for analysis consists of 281,779 single-family detached homes. For each categorical feature, binary encoding is used to create $k-1$ predictors, where k is the number of categories. Two additional predictors, squared average monthly temperature and squared average monthly humidity, are included in the linear regression model to account for any quadratic relationship between temperature/humidity and energy consumption.

4.4. Training and test sets

Training and test data sets are mutually exclusive in terms of both households and time (month and year), which provides a true out-of-sample prediction. Training sets are constructed using monthly household energy consumption for the 36 months from January 2010 to December 2012, and test sets are constructed using a separate set of households monthly energy consumption for the 12 months in 2013. To induce randomness in the modeling process, 10,000 homes are randomly sampled with the training set consisting of 7000 homes and the remaining 3000 for the test set. This sampling process is repeated 25 times with average performance metrics along with the corresponding standard errors reported. One caveat, is that actual average monthly temperature and humidity values are used in both the training and test data. Hence, results are given assuming full knowledge of future monthly temperature and humidity. Although monthly temperature and humidity cannot be predicted with absolute certainty, reporting results using actual values shows the ability of model performance assuming reliable monthly climate forecasts. To demonstrate the variation in model performance under uncertainty in climate forecasts, validation results are also reported using historical temperature and humidity values in the test data. Historical temperature and humidity values are obtained using the prior 30 years (1983–2012). For each test set, 30 modified test sets are constructed where each test set contains a prior years average monthly temperature and humidity values. Results using historical climate data are reported across all 750 modified test sets.

5. Results and discussion

5.1. Model selection

Model selection plots are useful for assessing model performance over model complexity, which aid in determining the most

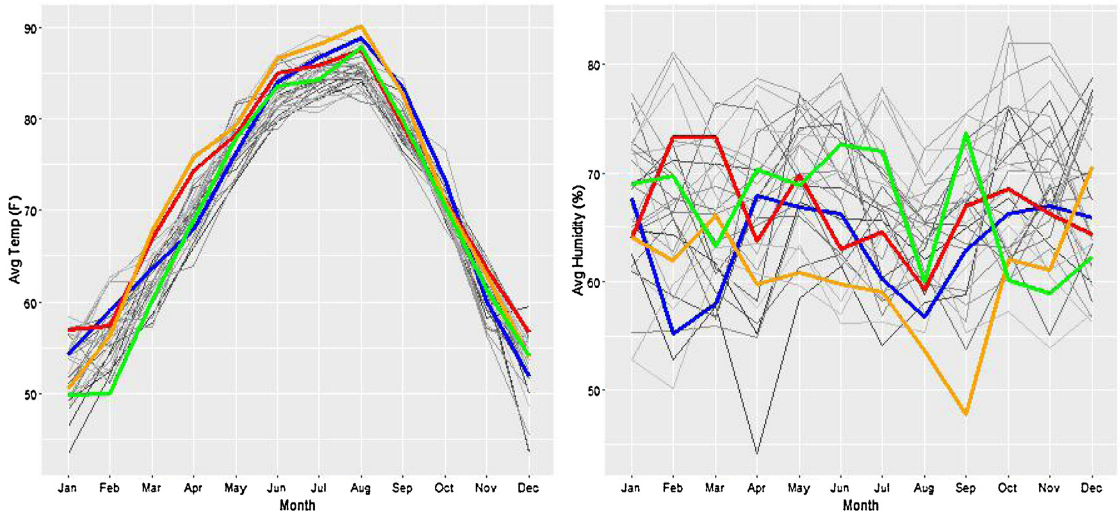

Fig. 4. Average monthly temperature (F) and humidity (%) 1983–2013, 2010 (green), 2011 (orange), 2012 (red), 2013 (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

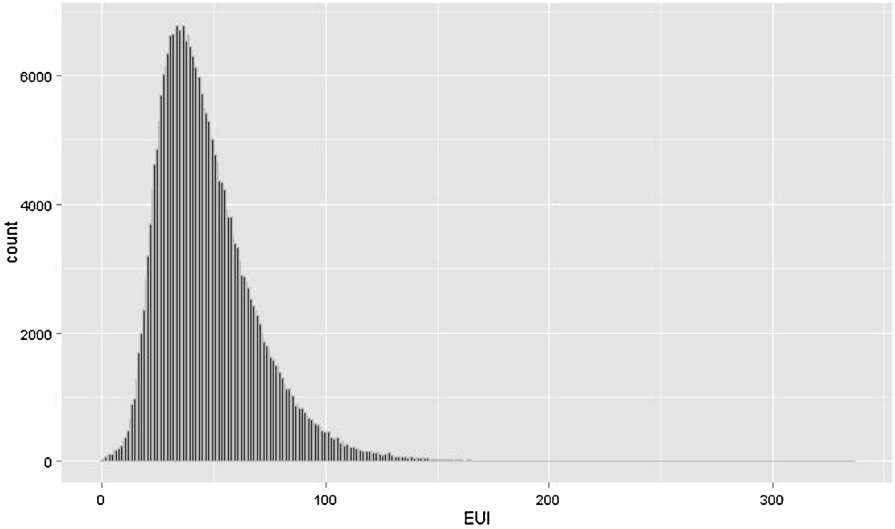

Fig. 5. Histogram of average EUI (kBtu/sf/year) across all homes with non-zero energy consumption for 2010–2013.

generalizable model for prediction and also serve to visualize the benefit of each additional level of complexity. As described in Section 3, model selection is conducted for all models. Linear Regression and MARS have similar techniques. For linear regression the full model without interaction terms is built and predictors are recursively removed based on the predictors that reduce the 10–fold cross-validated R^2 the least. For MARS each predictor and all pairwise interactions are considered for model building with the “best” model selected using 10-fold generalized cross-validated R^2. Generalized R^2 is based on the GCV statistic and has similar interpretation as R^2. Since model selection for regression trees does not perform feature elimination, model selection is illustrated by plotting 10-fold cross-validated R^2 versus number of splits. Fig. 6 shows the model selection plots for each model averaged across all training sets. Since the “best” linear regression model has no measurable difference from the full model and a similar number of predictors, the full model is selected as the “best” model. For MARS, the “best” model resulted with 29 terms (including the intercept) and 11 predictors, on average across all training sets. The “best” regression tree model resulted in a tree with 1706 splits on average. For linear regression, Fig. 6 shows that R^2 increases monotonically as the number of predictors increases; however, 95% off all the increase in R^2 from the intercept model occurs using the top 5 predictors, which represents the original features: size of living area, average monthly temperature, and fuel type. Similarly for MARS, 95% of all the gain in R^2 is generated from the top 10 terms (w/intercept) which include the features: size of living area, average monthly temperature, fuel type, vintage, and pool ownership. For regression trees 95% the gain in model performance is within the first 500 splits of the tree while 85% of the gain is captured by the first 100 splits. These plots suggest that models with much lower complexity (i.e., models with fewer predictors/terms /splits) are expected to be within 5% as accurate as the “best” models.

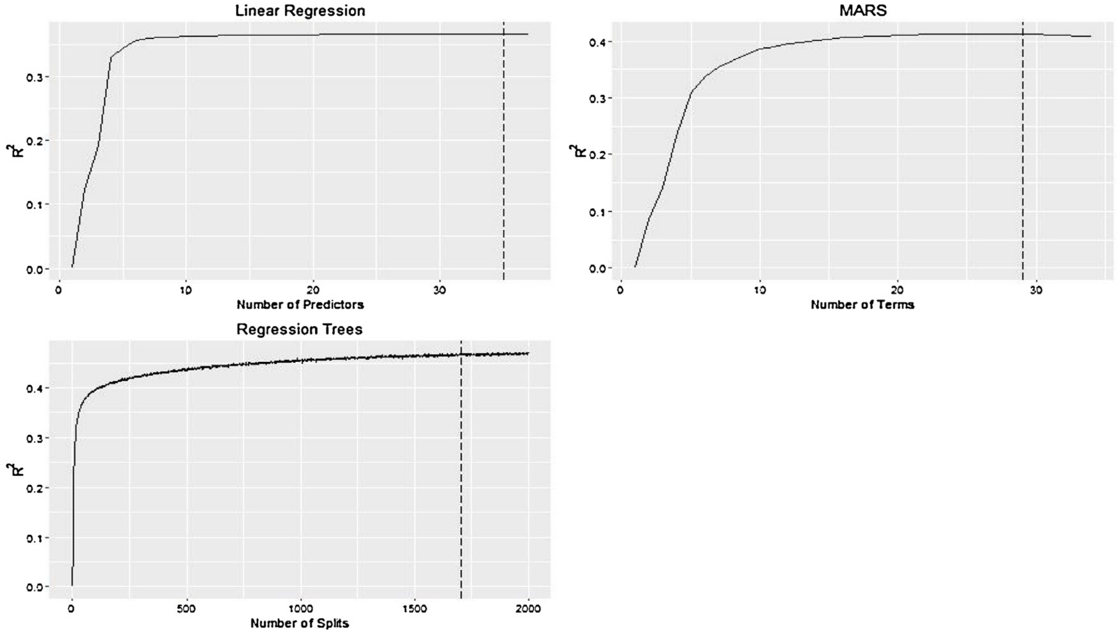

Fig. 6. Model selection plots for each model.

5.2. Variable importance

To understand the predictive benefit each feature adds to the model, variable importance statistics are computed. Since the importance of each predictor is a function of the model, different variable importance statistics are employed for each regression technique. For linear regression, the variable importance measure is the absolute value of the *t*-statistic corresponding to the coefficient estimate for each predictor. For MARS, the reduction in GCV is computed for each term added to the model. These reductions are summed for each corresponding predictor to get a variable importance measure. Predictors that are not included in the "best" model are given a variable importance of zero. For regression trees, the reduction in cross-validated error is measured at each split. This reduction is summed over each split corresponding to each predictor. Variable importance statistics for categorical features are averaged over each binary predictor, resulting in one variable importance statistic per original categorical feature (this is performed only for linear regression and MARS). Larger variable importance statistics imply more predictive value, given the model. Fig. 7 displays the bar plots of average variable importance for each model, over the 25 training sets. Across all models, size of living area, average monthly temperature, vintage, fuel type and pool ownership (except for regression trees) all significantly contribute to the predictability of monthly energy consumption. These findings should come as no surprise since they are well supported by other analyses [5–8,13]. However, there are meaningful differences in the magnitude and rank of each predictor between models. For regression trees, lot size is nearly as important as size of living area, exterior wall type is more important than average monthly temperature, and pool ownership and detached living area are ranked the least important. This is in contrast to the MARS model, where size of living area is more than 12 times as important as lot size, average monthly temperature ranks 2nd, pool ownership ranks 5th, and exterior wall type is the 9th most important feature. Interestingly, for linear regression exterior wall type is the least predictive feature unlike MARS and regression trees. For MARS, there is a high contrast between relevant and irrelevant features since variable importance is zero for predictors not included in the "best" model. From Fig. 7, the variable importance plot of MARS shows the following features have variable importance of zero or near zero: spa ownership, number of stories, number of half bathrooms, number of rooms, number of full bathrooms, attached garage, attached porch, roof covering type, fireplace, construction style and average monthly humidity. However, these features are not necessarily irrelevant predictors. A likely explanation is that some features provide similar information as other more important features, and hence are being masked. For instance, number of rooms is correlated with size of living area and number of bedrooms at 0.79 and 0.68 respectively, while number of stories has a 0.59 correlation with size of living area. In addition, fireplace is correlated with size of living area and vintage at 0.42 and 0.47 respectively, meaning that newer and larger homes are more likely to have fireplaces. While some features are masked, others seem to be of little to no importance across all models, such as: roof covering type, attached porch, number of half bathrooms, construction style and average monthly humidity.

5.3. Validation results

To assess the validity of each model at predicting future energy consumption, validation results are reported across all test sets. Overall performance metrics are measured using median absolute

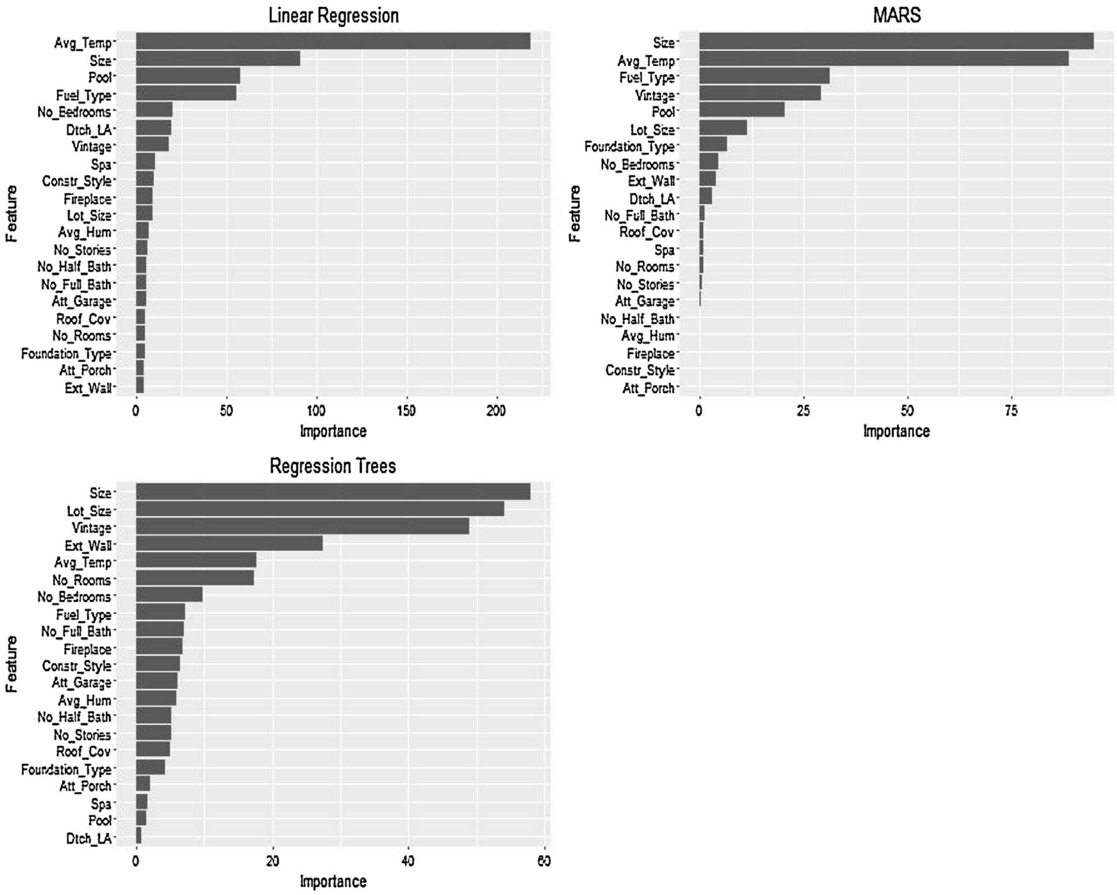

Fig. 7. Variable importance bar plots for each model.

Table 2
Summary of overall performance by model using actual climate data.

Model	MdAPE	RMSE	R^2
LR	0.2720 ± 0.0084	99.803 ± 3.057	0.4200 ± 0.0210
RT	0.2797 ± 0.0072	100.435 ± 3.441	0.4127 ± 0.0230
MARS	0.2606 ± 0.0071	94.286 ± 3.238	0.4824 ± 0.0198

percent error (MdAPE), root mean squared error (RMSE), and R squared (R^2). The formula for each performance metric is written as:

$$\text{MdAPE} = median\left\{\left|\frac{y_i - \hat{f}(x_i)}{y_i}\right|;\quad i = 1, \ldots, m\right\} \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}[y_i - \hat{f}(x_i)]^2} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}[y_i - \hat{f}(x_i)]^2}{\sum_{i=1}^{m}[y_i - \bar{y}]^2} \tag{7}$$

where $\hat{f}$ represents the predicted response and m is the number of observations in the test set. All performance metrics are calculated by comparing actual average daily energy (kBtu/day) to predicted average daily energy (kBtu/day); therefore, all predicted log response values are transformed to the original scale, which is average daily energy (kBtu/day).

5.3.1. Results at the household level

Tables 2 and 3 summarize the overall performance by model using actual and historical climate data.

From Table 2, MARS has significantly better performance across all metrics than both linear regression and regression trees, with MdAPE of 26%, RMSE of 94.286,and R^2 of 48.24. This implies that MARS model can predict future monthly energy consumption with about 74% accuracy for the median household, and within 94 kBtu/day on average (note: across all months the average household in 2013 consumed 215 kBtu/day). The R^2 value of 0.4824 implies that 48% of the variance in monthly household energy consumption can be accounted for using the building characteristics and average monthly temperature based on the MARS model. Regression trees slightly underperform linear regression, despite regression trees being capable of modeling more complex relationships. Results from Table 2 assume future average monthly temperature/humidity are fully known. Using historical climate data, Table 3 shows that performance across all models is statistically equivalent for the median household compared to the performance using actual monthly temperature and humidity. However, for the average household, there is a decay in performance for all models, on the order of 5% to 7% for RMSE and 12% to 14% relative to R^2. While there is a measurable loss in accuracy on average, all models are robust to uncertainty in climate forecasts measured by the median household.

A more detailed understanding of model performance is seen by measuring the accuracy of each model by fuel type. From Table 4, there are substantial differences in the predictability of monthly energy consumption between fuel types. Notably, households with fuel type 2 and 4 are the least predictable in terms of R^2 and MdAPE respectively, and households with fuel types 1, 3, and 5 have similar predictability in terms of R^2 and MdAPE. The largest contrast in performance between fuel types is seen when comparing RMSE for all-electric homes (fuel type 1) versus dual-fuel homes (fuel types 2, 3, 4,and 5). Though RMSE is not a standardized measure to compare predictability between fuel types, it does show that monthly household energy consumption can be predicted within 71.5 kBtu/day for all-electric homes versus 92–107.5 kBtu/day for dual-fuel homes, using MARS. This fact should be taken into account given that all-electric homes consume an average of 173.4 kBtu/day in 2013, and dual-fuel homes consume 214.6, 250, 175, and 235.4 kBtu/day in 2013 for fuel types 2, 3, 4, and 5 respectively.

Table 3
Summary of overall performance by model using historical climate data.

Model	MdAPE	RMSE	R^2
LR	0.2717 ± 0.0086	105.050 ± 5.127	0.3674 ± 0.0598
RT	0.2795 ± 0.0070	106.161 ± 5.481	0.3546 ± 0.0564
MARS	0.2596 ± 0.0067	101.232 ± 5.765	0.4123 ± 0.0648

Table 4
Summary of overall performance by fuel type and model using actual climate data.

Model	Fuel type	MdAPE	RMSE	R^2
LR	1	0.2487 ± 0.0054	73.098 ± 2.960	0.4741 ± 0.0363
RT	1	0.2655 ± 0.0060	78.717 ± 6.345	0.3906 ± 0.0715
MARS	1	0.2407 ± 0.0055	71.494 ± 3.098	0.4972 ± 0.0311
LR	2	0.2670 ± 0.0130	102.501 ± 14.301	0.2863 ± 0.1541
RT	2	0.2966 ± 0.0150	109.552 ± 7.102	0.1795 ± 0.1435
MARS	2	0.2671 ± 0.0162	100.559 ± 10.009	0.3165 ± 0.0940
LR	3	0.2517 ± 0.0050	113.843 ± 5.567	0.4339 ± 0.0324
RT	3	0.2585 ± 0.0051	113.988 ± 7.808	0.4325 ± 0.0539
MARS	3	0.2388 ± 0.0044	107.363 ± 6.409	0.4962 ± 0.0436
LR	4	0.3193 ± 0.0125	102.313 ± 4.146	0.3379 ± 0.0359
RT	4	0.3373 ± 0.0156	100.542 ± 4.226	0.3608 ± 0.0320
MARS	4	0.2958 ± 0.0112	92.505 ± 3.573	0.4584 ± 0.0345
LR	5	0.2627 ± 0.0257	112.902 ± 12.697	0.3348 ± 0.1005
RT	5	0.2668 ± 0.0323	104.167 ± 12.723	0.4346 ± 0.0881
MARS	5	0.2342 ± 0.0249	95.913 ± 12.829	0.5230 ± 0.0675

Performance by fuel type using historical climate data is shown in Table 5. For most fuel types there is a small decrease in model performance for the median household when compared to model performance using actual climate data, but a considerable loss in accuracy for the average household. Using MARS, RMSE increases 3.3%, 3.3%, 7.8%, 6.9% and 15.7% while R^2 decreases 7.4%, 14.6%, 16.8%, 17.2%, and 33.3% for fuel type 1, 2, 3, 4, and 5 respectively when using historical versus actual monthly average temperature/humidity values.

5.3.2. Results at the aggregate level

Using predictive models for forecasting monthly energy consumption at the household level can also be applied to predicting

Table 5
Summary of overall performance by fuel type and model using historical climate data.

Model	Fuel type	MdAPE	RMSE	R^2
LR	1	0.2795 ± 0.0191	82.692 ± 8.653	0.320 ± 0.1512
RT	1	0.2716 ± 0.0073	80.188 ± 6.186	0.367 ± 0.0707
MARS	1	0.2512 ± 0.0064	74.089 ± 3.164	0.460 ± 0.0330
LR	2	0.2827 ± 0.0154	108.728 ± 15.496	0.194 ± 0.1954
RT	2	0.3060 ± 0.0154	112.686 ± 7.261	0.132 ± 0.1490
MARS	2	0.2766 ± 0.0134	103.837 ± 9.522	0.270 ± 0.0993
LR	3	0.2581 ± 0.0082	117.200 ± 6.635	0.399 ± 0.0537
RT	3	0.2713 ± 0.0090	119.979 ± 7.956	0.370 ± 0.0627
MARS	3	0.2545 ± 0.0090	115.728 ± 8.949	0.412 ± 0.0846
LR	4	0.3197 ± 0.0137	103.126 ± 5.276	0.326 ± 0.0559
RT	4	0.3459 ± 0.0160	104.050 ± 5.026	0.315 ± 0.0488
MARS	4	0.3128 ± 0.0132	98.895 ± 6.989	0.379 ± 0.0883
LR	5	0.2582 ± 0.0290	113.101 ± 13.369	0.330 ± 0.1181
RT	5	0.2810 ± 0.0327	110.273 ± 12.337	0.362 ± 0.1126
MARS	5	0.2638 ± 0.0274	111.016 ± 17.648	0.349 ± 0.2052

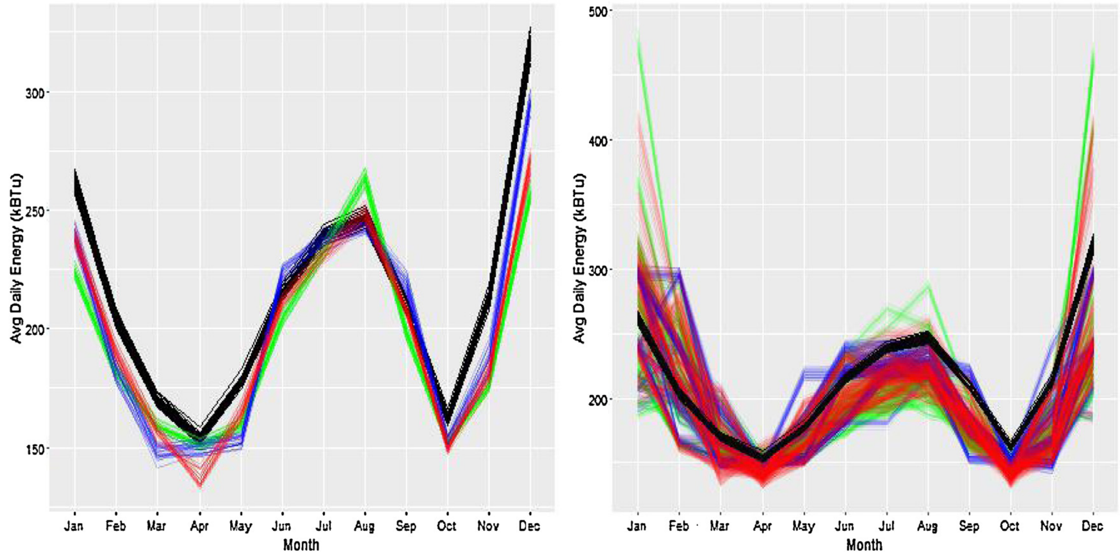

Fig. 8. Plot of mean predicted average daily energy (kBtu/day) for each month by model for all test sets using actual climate data (left) and historical climate data (right). Linear regression (green), regression trees (blue), MARS (red) and actual (black). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Summary of aggregate monthly performance by model using actual climate data.

Model	MdAPE	RMSE	R^2
LR	0.0692 ± 0.0040	25.743 ± 1.097	0.6784 ± 0.0254
RT	0.0698 ± 0.0054	18.277 ± 1.156	0.8373 ± 0.0216
MARS	0.0627 ± 0.0079	19.831 ± 1.187	0.8088 ± 0.0223

future total monthly energy for groups of households. To do this, the mean predicted average daily energy (kBtu/day) is compared to the mean actual average daily energy (kBtu) for 3000 homes in each test set. Performance is reported for mean predicted monthly energy across all homes in the test data for each model using actual and historical climate data. The average performance metrics along with their respective standard errors are shown in Tables 5 and 6. Regression trees significantly outperform both MARS and linear regression, in terms of RMSE and R^2 using actual and historical monthly temperature and humidity, although MARS is nearly as accurate as regression trees using actual climate data. From Table 6, total monthly energy is predicted for large groups of homes with a median accuracy of about 93%, and within 18.28 kBtu/day per household using regression trees. Since the average single-family detached home uses approximately 215 kBtu/day, this means total monthly energy is predicted with 91.5% accuracy on average, assuming accurate climate forecasts. Table 7 shows a sizable loss in accuracy for all models when uncertainty is introduced in climate forecasts. The RMSE nearly doubles for regression trees and MARS, and R^2 decreases by more than 50%. Predicted total monthly energy is expected to be within 32.61 kBtu/day per household of the actual total monthly energy consumption (i.e., approximately 85% accurate on average), based on regression trees model. An interesting note is that regression trees significantly underperform MARS at predicting monthly energy consumption at the household level, but outperform when predicting for large groups of homes. A property of regression trees is that the predicted value of a new observation, x, is simply the sample mean of the response values in the corresponding partition x belongs to from the training set. This implies that regression trees cannot extrapolate beyond the training space, unlike linear regression and MARS. This may be a reasonable explanation of why regression trees provide better aggregate predictions.

Table 7
Summary of aggregate monthly performance by model using historical climate data.

Model	MdAPE	RMSE	R^2
LR	0.1038 ± 0.0307	39.652 ± 9.181	0.197 ± 0.4067
RT	0.0897 ± 0.0227	32.610 ± 6.330	0.465 ± 0.1953
MARS	0.1040 ± 0.0206	35.227 ± 7.049	0.374 ± 0.2565

To visualize the performance of each model by month, Fig. 8 shows the monthly predicted average daily energy (kBtu/day per household) for each model versus the true average daily energy (kBtu/day per household) using actual climate data (left panel) and historical climate data (right panel). Fig. 8 illustrates the variability in predicted monthly energy consumption when there is uncertainty in average monthly temperatures. During the winter months, December, January, and February, all model predictions have a larger variance compared to fall, spring and summer months, which is a reasonable outcome since total monthly energy consumption is more variable during the winter compared to other seasons, as seen in Fig. 1. From the data in this study, December, January, and February have the largest variability in terms of average daily energy consumption (kBtu/day) with standard deviations of 170.1, 185.2, and 165.6 respectively. The next most variable month is August with a standard deviation value of 121.7. Another observation from Fig. 8 is that most predictions from June 2013 to December 2013 are substantially below actual energy consumption (black lines), using historical climate data. This phenomena is due to the fact that average monthly temperatures for the second half of 2013 are abnormally high for cooling months and abnormally low for heating months. The plot of historical average monthly temperatures from Fig. 3 supports this claim, showing that average monthly temperature for June 2013 to October 2013 are above historical averages while November and December 2013 are below. These findings indicate, that predicting future aggregate energy consumption is adversely affected when monthly temperature predictions significantly deviate from the actual values, but models tend to be robust to small inaccuracies in monthly temperature predictions.

6. Conclusion

Predicting future monthly energy consumption has wide implications for institutions that rely on accurate forecasts of residential energy. For example, utilities can incorporate the data-driven models presented in this paper to project power generation needs, prioritize investment projects, and simulate energy demand under changes in building characteristics and climate. Additionally, residential developers along with environmental agencies can improve the estimates of building energy demand and better assess the impact to carbon emissions. Reliable forecasting of energy consumption also helps end-users estimate the energy effects of building upgrades and could be used to educate consumers on how to reduce their energy consumption.

Methods presented in this study show that future monthly energy consumption for households outside of the training set can reliably be predicted using building characteristics and climate data. Using accurate monthly temperature forecasts, results show that monthly energy consumption at the household level can be predicted with 74% accuracy for the median household, and within 94 kBtu/day on average using MARS model, while forecasts for aggregate monthly energy consumption show a nearly 94% median accuracy and are within 18.3 kBtu/day per household using regression trees. When monthly temperature forecasts are uncertain, model performance is robust for the median household, but prediction error at the household level significantly increases to 101 kBtu/day and 32.6 kBtu/day per household at the aggregate level. Examining model performance by fuel type shows a meaningful difference in model performance between fuel types, specifically between all-electric and dual-fuel homes.

Model selection and variable importance plots illustrate that size of living area, vintage of home, fuel type, pool ownership, and average monthly temperature provide the most predictive value and other features such as lot size, exterior wall type, foundation type, and detached living area provide marginal predictive value. Other features including number of bedrooms, number of bathrooms, number of half bathrooms, number of stories, and presence of a fireplace while meaningful predictors provide similar information as other more important features and hence are mostly likely masked during model selection.

While the results from this study are encouraging important observations are worth highlighting to improve the process of predicting future residential energy consumption. For one, inaccurate climate inputs can significantly affect the accuracy of predicted residential energy usage, especially at the aggregate level. This implies energy predictions should be computed over a range of climate values and utilizing a diverse set of models. Second, residential energy consumption during winter months are more variable than other seasons for this region of study. This may be a result of the variable temperatures during the winter months. Abnormally cool winter months seem to result in a larger marginal increase in residential energy consumption than abnormally warm summer months. Understanding the behavioral response to cooler winter temperatures in “mild” winter climates needs more investigation. Third, climate is not uniform within a given region (e.g., county/city) or even between nearby households. Thus, collecting weather data from different locations within the region of study provides a more representative climate sample than a sole weather station. Currently there does not exist a network of weather sensors collecting and sharing the data necessary for this analysis. However, such a network may be available in the future, given developments in technology, reduction in sensor costs, development of new applications, and a greater understanding from stakeholders of the potential value of such information. Lastly, a small portion of households consumes little to no energy while others have abnormally high usage when comparing energy utilized by similar single-family detached homes. While understanding the nature of this abnormal behavior is outside the scope of this study, the impact of behavioral variables on total energy consumption is not negligible; therefore, accurately accounting for these outlying energy consumers is necessary.

There are many directions for future research, including: forecasting long term trends in residential energy based on changes in climate and building stock, incorporating historical household energy consumption to build better predictive models, and using spatial data to understand the effects of location on residential energy demand.

Acknowledgements

This project and the preparation of this study were funded in part by monies provided by CPS Energy through an agreement with the University of Texas at San Antonio. Special thanks to Rahul Nair, Vidal Ramirez and Carlos Contreras for their help processing and pairing data required for this manuscript.

References

[1] P. de Wilde, The gap between predicted and measured energy performance of buildings: a framework for investigation, Autom. Constr. 41 (2014) 40–49, http://dx.doi.org/10.1016/j.autcon.2014.02.009.

[2] T. Peffer, W. Burke, D. Auslander, Response: modeling the wide variability of residental energy consumption, in: Proc. of ACEEE Summer Study on Energy Efficiency in Buildings, 2010.

[3] U.S. Energy Information Administration, Drivers of U.S. Household Energy Consumption 1980–2009, 2015, URL https://www.eia.gov/analysis/studies/buildings/households/pdf/drivers_hhec.pdf.

[4] L. Swan, V. Ugursal, Modeling of end-use energy consumption in the residential sector: a review of modeling techniques, Renew. Sustain. Energy Rev. 13 (8) (2009) 1819–1835, http://dx.doi.org/10.1016/j.rser.2008.09.033.

[5] J. Gomez, A. Elnakat, M. Wright, J. Keener, Analysis of the energy index as a benchmarking indicator of potential energy savings in the san antonio, Texas single-family residential sector, Energy Effic. 8 (3) (2015) 577–593, http://dx.doi.org/10.1007/s12053-014-9310-6.

[6] J. Min, Z. Hausfather, O.F. Lin, A high-resolution statistical model of residential energy end use characteristics for the united states, J. Ind. Ecol. 14 (5) (2010) 791–807, http://dx.doi.org/10.1111/j.1530-9290.2010.00279.x.

[7] K. Steemers, G.Y. Yun, Household energy consumption: a study of the role of occupants, Build. Res. Inform. 37 (5–6) (2009) 625–637, http://dx.doi.org/10.1080/09613210903186661.

[8] G.Y. Yun, K. Steemers, Behavioural, physical and socio-economic factors in household cooling energy consumption, Appl. Energy 88 (6) (2011) 2191–2200, http://dx.doi.org/10.1016/j.apenergy.2011.01.010.

[9] C. Valenzuela, A. Valencia, S. White, J. Jordan, S. Cano, J. Keating, J. Nagorski, L. Potter, An analysis of monthly household energy consumption among single-family residences in Texas 2010, Energy Policy 69 (2014) 263–272, http://dx.doi.org/10.1016/j.enpol.2013.12.009.

[10] Z. Yu, F. Haghighat, B. Fung, H. Yoshino, A decision tree method for building energy demand modeling, Energy Build. 42 (10) (2010) 1637–1646, http://dx.doi.org/10.1016/j.enbuild.2010.04.006.

[11] R. Edwards, J. New, L. Parker, Predicting future hourly residential electrical consumption: a machine learning case study, Energy Build. 49 (2012) 591–603, http://dx.doi.org/10.1016/j.enbuild.2012.03.010.

[12] E. Hosgor, P. Fischbeck, Predicting residential energy and water demand using publicly available data, Energy Convers. Manage. 101 (2015) 106–117, http://dx.doi.org/10.1016/j.enconman.2015.04.081.

[13] J. Kolter, J.A. Ferreira, A large-scale study on predicting and contextualizing building energy usage, in: Proc. Twenty-fifth AAAI Conference on Artificial Intelligence, 2011.

[14] B. Dong, C. Cao, L.E. Siew, Applying support vector machines to predict building energy consumption in tropical region, Energy Build. (2005) 545–553, http://dx.doi.org/10.1016/j.enbuild.2004.09.009.

[15] T. Catalina, J. Virgone, E. Blanco, Development of validation and regression models to predict monthly demand for residential buildings, Energy Build. 40 (2008) 1825–1832, http://dx.doi.org/10.1016/j.enbuild.2008.04.001.

[16] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Chapman and Hall, 1984.

[17] T.M. Therneau, B. Atkinson, B. Ripley, rpart. Recursive Partitioning, R Package Version 4.1.9, 2015, URL https://CRAN.R-project.org/package=rpart.

[18] T. Hastie, R. Tibshirani, Generalized additive models, Stat. Sci. 1 (3) (1986) 297–310.

[19] J.H. Friedman, Multivariate adaptive regression splines (with discussion), Ann. Stat. 19 (1991) 1–141.

[20] M.Y. Cheng, M.T. Cao, Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines, Appl. Soft Comput. 22 (2014) 178–188, http://dx.doi.org/10.1016/j.asoc.2014.05.015.

[21] H. Friedman, C. Roosen, An introduction to multivariate adaptive regression splines, Stat. Methods Med. Res. 4 (1995) 197–217, http://dx.doi.org/10.1177/096228029500400303.

[22] S. Milborrow, Derived from mda:mars by, T. Hastie, R. Tibshirani, earth: Multivariate Adaptive Regression Splines, R Package Version 4.4.4, 2016, URL https://CRAN.R-project.org/package=earth.