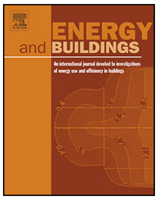
Energy and Buildings 128 (2016) 1–11

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