Building and Environment 138 (2018) 194–206

Contents lists available at ScienceDirect

Building and Environment

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

# Evaluation of the causes and impact of outliers on residential building energy use prediction using inverse modeling

Huyen Do[a,b], Kristen S. Cetin[a,*]

[a] Department of Civil, Construction and Environmental Engineering, Iowa State University, Ames, IA, United States
[b] Faculty of Project Management, The University of Danang - University of Science and Technology, Danang, Viet Nam

ARTICLE INFO

*Keywords:*
Inverse modeling
Energy use
Residential buildings
Outliers
Building performance

ABSTRACT

Inverse modeling techniques are often used to predict the performance and energy use of buildings. Residential energy use is generally highly dependent on occupant behavior; this can limit a model's accuracy due to the presence of outliers. There has been limited data available to determine the cause of and evaluate the impact of such outliers on model performance, and thus limited guidance on how best to address this in model development. Thus the main objective of this work is to link the use of outlier detection methods to the causes of anomalies in energy use data, and to the determination of whether or not to remove an identified outlier to improve an inverse model's performance. A dataset of 128 U.S. residential buildings with highly-granular, disaggregated energy data is investigated. Using monthly data, change-point modeling was determined to be the best method to predict consumption. Three methods then are used to identify outliers in the data, and the cause and impact of these outliers is evaluated. Approximately 19% of the homes had an outlier. Using the disaggregate data, the causes were found to mostly be due to variations in occupant-dependent use of large appliances, lighting, and electronics. In 20% of homes with outliers, the removal of the outlier improved model performance, in particular all outliers identified with both the standard deviation and quartile methods, or all three methods. These two combinations of outlier detection methods are thus recommended for use in improving the prediction capabilities of inverse change point models.

## 1. Introduction

In recent years, the energy consumption in buildings has continued to increase, accounting for approximately 40% of worldwide energy consumption [1]. In the U.S. in 2015, energy use in residential and commercial buildings represented approximately 40% of total energy consumption [2]. Building energy consumption accounts for one-fifth of total global energy use [3]. In addition, the total building energy use worldwide is forecasted to increase an average of 1.5% per year from 2012 to 2040 [3]. This increasing energy utilization in buildings strongly affects the environment and climate. The energy consumption in buildings currently accounts for approximately one-third of the current greenhouse gas (GHGs) emissions worldwide [1], and 12% in the U.S [4]. Thus, the identification and application of methodologies to decrease the buildings energy and electricity demands is important given global energy challenges as well as the impending consequences of climate change [5].

Energy efficiency improvements to the existing and future building stock helps accomplish these energy reductions. Several of these upgrades include improving the design and construction of new buildings, and retrofitting existing buildings with higher-efficiency equipment, higher-efficiency materials for the building envelope, and more intelligent and efficient controls and control strategies [6–8]. These methods do not require human intervention and are not dependent on occupant interaction with the building to save energy. However, much of a commercial or particularly residential building's energy use is also dependent on occupant behavior. Recent studies have found that 19% of energy use can be explained by variations in occupants' use of the building [9,10], and that when end-uses such as plug loads and appliances are dependent on occupant interaction they are up to 10 times more variable over time than those that are not [11]. Occupant behavior can make energy use more unpredictable, but it is also possible to influence occupant-dependent energy use as a method of energy conservation.

Behavioral energy efficiency is a generally lower-cost method to achieve energy savings [12,13]. This method's purpose it to change occupant energy-related behaviors [14] by providing feedback to customers through either a direct (real-time or near real-time) or indirect

* Corresponding author. Tel.: +1 515 294 8180.
*E-mail addresses:* huyendt@dut.udn.vn, huyendo@iastate.edu (H. Do), kcetin@iastate.edu (K.S. Cetin).

https://doi.org/10.1016/j.buildenv.2018.04.039
Received 4 January 2018; Received in revised form 29 March 2018; Accepted 27 April 2018
Available online 30 April 2018

(post-consumption) method. Previous studies have found that these methods can achieve on average from 2 to 7% energy savings depending on the frequency and type of energy information provided [15], or an average of approximately 4% savings for real-time feedback programs and 2% for enhanced billing strategies [16]. However, real-time feedback strategies are not possible for many households, since for approximately half of households in the U.S., monthly energy use is the only energy consumption information available [17,18]. For these homes, indirect feedback can provide information to aid in behavior-motivated energy savings.

The feedback provided to residential customers most commonly includes information such as whole-home energy use, disaggregated end uses, future forecasted energy use, and/or comparison with neighboring homes' performance [15,19]. This information is typically determined through the development and use of data-driven models trained using historic energy data. As a result of such information, customers are better informed about their energy behaviors, and also better understand through recommendations developed through these insights, what energy savings they can achieve. These insights drive energy saving behaviors [20]. In the use of data-driven, inverse models [21,22] for energy efficiency behavioral changes, the accuracy of such models is highly important to ensure the homeowner trusts the results and predictions of such model enough to invest time and effort into efficiency upgrades. This method has advantages over calibrated building energy simulation methods, including limiting the need for detailed building information, and the ability to provide near-instantaneous results [23].

Inverse modeling techniques typically use outdoor weather data, including temperature, wind speed, humidity, and solar radiation as the main predictor(s) of the energy use of a building [24]. As weather significantly influences residential energy use, often the use of weather data as the input into these models is sufficient to provide a reliable model. A variety of inverse modeling methods have been applied for the prediction of building energy use in recent literature [25]. Change-point models [23,24] typical use outdoor dry-bulb temperature as the independent variable to predict building energy use using a combination of regression analysis methods and the determination of a balance point between trends in energy use trends by season. Artificial neural networks (ANN) are a supervised machine learning method which includes input layers, hidden layers, and output layers to predict the energy consumption, often applied to more frequent datasets such as daily, hourly, or sub-hourly data [26–28]. Similarly, genetic programming uses an evolutionary algorithm to automatically compute data and make the prediction from biological process; this method has been applied to predict residential HVAC use or commercial building energy [29,30]. Probabilistic graphic models such as Bayesian networks [31,32], and Gaussian mixture models [23,33,34] with multivariate nonlinear regression function have been shown to be able to predict the energy consumption in the building using monthly, daily or hourly frequency data. Recent research has also utilized other models including support vector machines (SVR) [30,35,36], hybrid model predictive control schemes [37], or occupant behavior models [38–42] to predict residential building energy use. In these types of data-driven models, there are many factors that may affect the accuracy and computation of such models. In most models, the data frequency such as the monthly, daily, hourly, or sub-hourly data is a strong factor that has direct influence on the performance of models [25]. The requirement of significant input training data in ANN method, genetic programming, and probabilistic graphic methods increases the associated computational demand [23]. The presence of outlier data points also impacts the fitness of these models [23,25].

However, when applying these inverse modeling techniques, particularly for residential buildings, a variety of uncontrollable factors, including occupant behavior, can have a strong influence on building energy performance, and can also result in significant variations in use. For example, Kim et al. (2015) [43] observed that the energy use of a residential building was extremely low during the summer season, then through contact with the owner, determined they were on vacation during that period. Therefore, in such cases, the energy consumption prediction using the inverse model might not align with the actual performance. The accuracy of inverse models, thus, is limited in these households if such outlier behaviors occur, as the occurrence of an outlier can influence the model prediction. In addition, in most cases it is not possible to make contact directly with a homeowner as suggested in Kim et al. [43], for better understanding the actual reasons for the energy use outlier in that month and it resulting treatment.

It is generally recommended in measurement and verifications (M&V) procedures to identify and remove outliers from datasets in the development of energy use predictions [44]. However, the decision to remove an outlier is often also dependent on the judgement of the modeler to determine, and typically requires justification beyond solely statically reasons as to why a particular data point should be removed. Without additional information to understand the causes of such outliers in residential energy use datasets, the removal of data is challenging to justify. If the decision to keep or remove a data point outlier is made through incorrect assumptions or justification, this may negatively influence the model accuracy. Therefore, it is necessary to further study the occurrence of energy use outliers in the inverse modeling techniques, including determining the possible reasons for the occurrence of these outliers and their influence of the model performance. This can help in developing a method for how to identity, assess, and treat these outliers in inverse modeling, and a stronger understanding of why such outliers occur in residential energy use. This is accomplished with the ultimate goal of better prediction of energy use in residential buildings, the results of which can also drive energy saving behaviors.

In this research a large dataset of residential buildings with highly-granular, disaggregated energy use data is investigated to determine the existence of outliers in inverse models developed from monthly energy use data, the most common type of energy data available for residential buildings in the U.S. First, three different methods including the standard deviation method, quartile method, and Grubbs' test are applied for outlier recognition in the developed inverse change-point models of residential building energy use data. Then, the specific reasons for the occurrence of the outliers are investigated using highly detailed and disaggregated data in these homes. Next, the impact of keeping or removing outliers these outliers on the performance of the inverse models are evaluated to ultimately determine the best methods for better prediction of energy use in the inverse models. Finally, the limitations, conclusions and future work are also discussed.

## 2. Methodology

In this study, the methodology is divided into two main stages. The first stage (Fig. 1) develops an inverse model and detects the presence of outliers in the inverse model. The second stage (Fig. 2) determines the causes of the outliers and whether it is recommended that these outliers be included in the final model based on their impact on the model performance with in- and out-of-sample data.

### 2.1. Outlier detection methodology

#### 2.1.1. Step 1 - data filtering and quality control

This step involves the collection, processing, quality control, and characterization of building energy use data and weather data. Each house in the dataset must have at least one year of data but preferably more than one year. This is needed in order to cover the energy performance of the building throughout all seasons with some additional data for use as out-of-sample data to test the model performance, and have over a minimum threshold percent of total data used to create the monthly energy use data dataset. The focus of this work is on outliers in monthly data, thus when considering houses with monthly data values,