Energy & Buildings 172 (2018) 116–124

Contents lists available at ScienceDirect

Energy & Buildings

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

Inverse energy model development via high-dimensional data analysis and sub-metering priority in building data monitoring

Zhao Chen[a], James Freihaut[b], Bo Lin[c,*], Christina Dan Wang[d]

[a] School of Data Science, Fudan University, China and Department of Statistics, the Pennsylvania State University, United States
[b] Department of Architectural Engineering, the Pennsylvania State University, United States
[c] SmithGroupJJR, United States
[d] New York University Shanghai, China and Department of Statistics, Columbia University, United States

ARTICLE INFO

Article history:
Received 27 December 2017
Revised 8 April 2018
Accepted 24 April 2018
Available online 3 May 2018

Keyword:
Big data
Building energy model
Office building
Inverse model
High dimensional data analysis
LASSO
Variable selection

ABSTRACT

In the US, building sector consumes approximately 41% of all U.S. primary energy and 70% of all generated electricity. Office building uses the largest percentage of primary and derived energy in the commercial buildings sector. Therefore, energy saving is the primary target of building renovation and equipment retrofit. To estimate energy savings after a renovation or equipment retrofit, measurement and verification (M&V) must be conducted based on energy use models or benchmark tools to track and assess savings. Although some methods and modeling tools are applied more often, there is no well-accepted model formulation methodology and some methods have high uncertainty.

This paper discusses a method to formulate the energy use model as an inverse regression model via an advanced and robust high-dimensional data analysis method. According to the inverse regression model, independent variables determined to be most influential to the energy consumption should be metered prior to building retrofit or audit. To select the most influential variables, this research considers a comprehensive set of potential key variables to avoid underspecified model issue. The models established by this method are observed to have great power in predicting the building energy consumption. More importantly, through the variable selection procedure, the method identifies key variables that should be monitored to continuously improve the building energy performance. A simulation study and a case study are presented to show the effectiveness of the new approach.

1. Introduction

In the era of "Big data", massive data are easily collected and stored by automatic systems and advanced digital equipment. The consequent revolution in data analysis technology has brought about new changes in scientific and industrial fields. Many successful examples showed that the application of big data analysis technology can effectively improve the original models or methods. In building industry, the big data analysis technology is also expected to contribute to improve the building energy use models. The energy use model is typically a statistical or thermal balance model that predicts building energy use. It is an important analytical tool in measurement and verification (M&V) and energy retrofit projects. These projects reduce operational energy expenditures and carbon footprint [1–4], and have given rise to a new sector in the building industry, the Energy Service Company (ESCO). With the advent of new sensors and advancement in information technology, high volume data about building information are obtained, causing troubles in the traditional way of modeling the energy use. In this paper, we proposed applying big data technology to develop building energy use model to predict building energy saving.

Energy savings from various energy conservation measures (ECMs) are calculated through the energy use model with data collected prior to and after retrofit. The building energy use model depends on lots of building variables and operating conditions, which are metered or recorded in the energy audit process. To avoid the model bias, engineers always attempt to consider as many variables as possible and compare different models. Different conclusions are derived based on different energy use models. For example, Kissock and Eger [5] showed that dry bulb temperature is more important than other possible environmental variables such as humidity, solar radiation. In addition to dry bulb temperature, many other variables have also been investigated in the literature. Sharp [6] indicated that floor area and temperature should be considered for energy variation. Chung et al. [19] included the building age, operational schedule, number of people, chiller type, and

* Corresponding author.
E-mail address: bo.lin@smithgroupjjr.com (B. Lin).

https://doi.org/10.1016/j.enbuild.2018.04.061

lighting equipment and control. Spyrou et al. [6]. applied a multivariate model with area of spaces, volume of sales (US dollar) and weather to estimate energy use. Bohdanowicz and Martinac [7] divided variables into several groups such as hot water use, services and facilities' operational factors, house laundry load and dining. Mathieu et al. [9] introduced time-of-week as a new variable in the demand response model that is effective in predicting hourly building energy use. As more data are collected, new variables are constantly introduced to establish energy use models.

However, with the increase of the number of variables, the uncertainty and complexity of the model are increasing due to the lack of adequate and high-quality observations. These two issues decrease the ability of interpretation and accuracy of prediction. For instance, only limited observations are collected because the time length of data measurement is insufficient due to project constraints. The limited number of observations has to cope with a considerably large number of possible variables. This causes the model fitting to be unstable and prone to be over-influenced by extreme conditions or rare events. Therefore, it is critical to select key variables from numerous measured variables.

The traditional methods to reduce the dimension of variables are principle component analysis (PCA) and variable selection. Among them, PCA has been widely applied in building energy modeling. Ruch et al. [8] adopted PCA to identify the relationship between outdoor conditions (dry-bulb temperature, solar radiation and humidity) and electricity consumption. Reddy and Claridge [10] compared PCA method with four-parameter linear model. More literature is referred to [11,12]. However, there are disadvantages in PCA. New variables, obtained by linear combination of original variables, usually do not have explicit explanation or interpretation of the data. Compared with PCA, variable selection can obtain a model with clear interpretation. However, variable selection is only suitable for the case with less candidate variables and sufficient sample size. In addition, traditional variable selection is time consuming and extremely susceptible to outliers and rare events

Solving the problems with aforementioned two methods involves the new technology called high-dimensional data analysis. High-dimensional data is one of the most important data types in "Big Data" field, which emphasizes that the number of observations is close to or even less than the number of variables under consideration. The lack of observations violates the basic condition of traditional data analysis, but inspires the development of high-dimensional approaches. The high-dimensional data analysis methods have been well studied in statistics during the last decade and widely applied in many fields. On one hand, it allows us to establish the model as fully as possible without the constraints on the number of observations. On the other hand, high-dimensional data analysis method develops data-driven model, which means that everything is determined by data with minimum assumptions. The fundamental problem in high-dimensional data analysis is how to reduce the dimension of variables and keep the most influential variables from a large number of candidates.

By applying high-dimensional analysis to building data, a data-driven, evidence - based energy use model will be constructed. The model statistically indicates the relationship between variables (for example ambient weather conditions) and output such as total energy consumption. High-dimensional data analysis method provides a robust, automatic methodology to determine the key variables. Meanwhile, it guarantees the efficiency and unbiasedness of the final model with the selected variables.

In this paper, the main purpose is to introduce high-dimensional data analysis to formulate the multivariate energy regression model and show the benefit with the new analysis. This proposed method can be applied to (1) identify critical variables that are statistically significant to overall building energy use, (2) develop robust energy prediction model in measurement and verification, and (3) make new sub-metering plan to measure identified key variables. Due to the abundance of high-dimensional data analysis methods, we illustrate the application and benefit especially with "least absolute shrinkage and selection operator" (LASSO) to simultaneously select the key variables and formulate the model for building energy use data. Other high-dimensional data analytic techniques can certainly by applied in place of LASSO according to different data types. The reason to pick LASSO as the example is that LASSO is more intuitive to understand and that some researchers have already been applying LASSO for different purposes in this field of studying energy. For example, Burling et al. [13] studied the incidence and impacts of energy efficiency project of schools in California. They explored a regression approach and a machine learning approach to predict energy consumptions with high frequency data. Jain et al. [14] applied LASSO method to determine the most important subset of variables for multi-family residential buildings and produced models to forecast energy consumption. In our research, we will focus on commercial buildings instead of residential. Clearly, their concerns are not the high-dimensional issue and they were studying different data types for different models. But it is encouraging to see that the current article is on the same page with the previous works on exploring the powerful new techniques in a different fashion and for different purposes.

The rest of the paper is organized as follows. In Section 2, we introduce the basic model and LASSO method. In Section 3, benchmark building case study is provided, which includes three parts: (1) building information is briefly discussed in Section 3.1; (2) input data preparation in Section 3.2; (3) results of case study. In Section 4, we test the method with an actual building data in Pennsylvania, U.S. In Section 5, we summarize the whole paper. Details of all variables are provided in the appendix.

2. Methodology

Modern high-performance buildings integrate many green sustainable strategies such as day-lighting and highly efficient systems to reduce energy use. The LEED certification and the Living Building Challenge developed tremendous needs for high-performance design. In addition to the new construction sector, the existing buildings also need performance improvement. Although about 12% of commercial buildings were built since 2003, most of the building sectors are fairly old [15]. More than half of the buildings were built before 1980, and the median age of buildings in 2012 was 32 years old [15]. The existing buildings have to rely on renovation to achieve energy and cost reduction. There are several main steps to conduct energy renovation in existing buildings and are briefly introduced in the following section.

2.1. Typical steps of energy renovation

In practice, the energy renovation involves 4 typical steps as follows. The first step is the building data auditing and collection: at this phase, building information such as space information and mechanical system conditions is collected by on-site visit or from building management systems (BMS) and drawings. After information collection, various meters are installed inside the building to record energy trends based on sub-system categories and space conditions. For example, building total energy use is of the most interest. Supply airflow, temperature, lighting energy and so on are metered as well. Typically, metered data at daily interval are accurate enough for analysis. Data are grouped into weekday data and weekend data because of distinct energy use patterns between the two categories. An energy model is developed at this step such