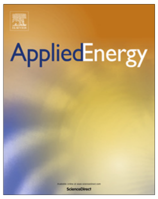
Contents lists available at ScienceDirect

Applied Energy

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



Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests

Jun Ma, Jack C.P. Cheng *

Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China

HIGHLIGHTS

- The influence of 171 different kinds of features was analyzed using Random Forests.
- Average energy use intensities of 1322 regions were set as the regression target.
- The model built by Random Forest has lower MSE than Lasso and SVM.
- An educational feature was found to be the most influential.
- The study not only identifies the influential features, but also matches the areas.

ARTICLE INFO

Article history:
Received 19 May 2016
Received in revised form 2 August 2016
Accepted 17 August 2016

Keywords:
Energy use intensity (EUI)
Multi-family residential buildings
Random Forests (RF)
Regional effect
Variable importance

ABSTRACT

Efficient and effective city planning in improving the energy performance of residential buildings requires a clear understanding of the influential features. Previous studies on modeling the relationships between influential features and the energy consumption have several gaps and limitations, such as the linear modeling methodology and insufficient consideration of particular features. This study therefore aims at investigating the influence of 171 possibly related features on the regional energy use intensity (EUI) of residential buildings using a non-linear regression algorithm, namely Random Forests (RF). The New York City (NYC) was focused on due to data availability. The 171 features covered seven different aspects, which are building, economy, education, environment, households, surrounding, and transportation. The average site EUI of the residential buildings in each Block Group (BG) was set as the dependent variable. The regression model was compared to the models using typical linear methods, such as Multiple Linear Regression and Lasso. The results show that the RF model achieved a lower mean square error. In addition, the top 20 influential features were identified based on the out-of-bag estimation in RF. Results show that less percentage of well-educated people, higher percentage of households heated by fuel oil, lower household income and more residential complaints per capita are correlated with higher average site EUI in NYC. Related suggestions on improving the energy performance in different regions are presented to the local government.

1. Introduction

The built environment accounts for around 40% of the total energy consumption in many developed countries [1]. Improving the energy performance of buildings is already a global target in both academia and industry [2]. Among the major components in building energy consumption, the residential sector is the largest part in many countries, such as the U.S. [3] and the U.K. [4] Therefore, studying the influential features on the energy use intensity of residential buildings is crucial in order to better understand the patterns and characteristics, and thus helps in policy making and in reducing the energy consumptions in the residential sector.

Various kinds of features have been studied in the past to investigate their influence on the energy consumption of residential buildings on the urban scale. The majority of the studied features can be grouped into three aspects, namely (1) the building aspect, (2) the household aspect, and (3) other aspects. The building aspect includes the features that describe the characteristics of residential buildings. Studies have shown that features like floor area, dwelling type, and dwelling age, are more important than other features in explaining the energy consumption [4,5]. For example, the floor area is concluded to be positively correlated with the energy consumption [2,4,5]. Different dwelling types, such as one family

* Corresponding author.
E-mail addresses: jmaae@ust.hk (J. Ma), cejcheng@ust.hk (J.C.P. Cheng).

http://dx.doi.org/10.1016/j.apenergy.2016.08.096
0306-2619/© 2016 Elsevier Ltd. All rights reserved.

house, multi-family residential buildings and mixed buildings, will have different energy estimation models, and they should be classified before regression [2,6]. The dwelling age generally reflects a negative correlation with the energy consumption [4,7]. However, most of these highly influential features cannot be easily changed through energy-efficiency retrofitting, and thus reveal limited help for actionable improvements [5]. Other features in the building aspect include material types and insulation level of walls, floors and windows. Although some studies reveal reasonable correlations between these features and energy consumption, their impacts were significantly low, and some even show nearly no impact [4,5,8].

The second aspect of influential features is household characteristics. Studies have shown that the variance of the energy consumption of similar buildings in the same location largely relied on the household characteristics [6,9]. Commonly seen features include household income, household size, density of the households, age of the householders, etc. Related research showed that building energy consumption generally increases with higher income [10–12], larger household size [5], and higher density of households [13]. The age of householders reveal different strengths of correlation in different studies [4,5], which may be resulted from the inconsistency and different quality of questionnaires used in surveys as Huebner et al. [5] suggested.

Besides the building aspect and household aspect, other features have also been studied to investigate their influence on the energy consumption of residential buildings. These features include, but are not limited to, climate [14], education [15], attitude of the occupants towards energy saving [5], water usage [16], culture [17], etc. However, except for climate and education, the majority of the features revealed limited contribution in estimating the energy consumption [14–17].

Although previous research has investigated many kinds of influential features on the energy consumption of residential buildings, there are still several gaps and limitations which need further studies to address. These can be summarized into the following four aspects. (1) Many influential features are non-linearly related to the energy consumption in the real world. For example, although the floor area appeared to be one of the most important estimators in many literatures [2,7,8,18], the correlation between the floor area and the building energy consumption never becomes ±1. However, the majority of the previous studies model the relationships based on linear regression methodologies, such as Multiple Linear Regression and Lasso (Least Absolute Shrinkage and Selection Operator) [4,5,16,19]. This limits the model performance and the discoveries. (2) The majority of the related studies use the whole building energy consumption as the regression target [4,5,18], and this may assign higher weights to the features that are correlated to the floor area, which is a non-changeable feature in existing buildings. As a result, some discovered relationships may imply limited actionable suggestions for city planning. (3) Most of the past research efforts use one individual building or dwelling as a case in training the regression model [4,5,15,16,18], so the discovered relationships or suggestions may fit more to individual buildings instead of a group of buildings. However, from the perspective of city planning and policy making, it is more useful to identify relationships or problems that have a group or regional effect, the study of which is currently lacking. (4) There are many features that may also affect building energy consumption but excluded in previous research. For example, some environmental features such as the surrounding vegetation may affect the heat island effect [14,20], and thus influence the energy consumption. The density of surrounding facilities such as shopping malls and subway stations may also affect the living pattern of occupants and thus affects the building energy consumption. Nevertheless, these features are rarely studied in previous research.

This study was therefore conducted to address the gaps mentioned above. The objectives are to use the non-linear based machine learning algorithm namely Random Forests to investigate the most influential features on the regional energy use intensity (EUI) of residential buildings, and to provide actionable suggestions on energy efficiency policy making. Due to data availability, this study focused on New York City (NYC), USA. Features ranging from the building aspect, the education aspect, the household aspect, the economy aspect, the environment aspect, the transportation aspect and the surrounding aspect, with a total number of 171 features, are included in this study. Details of the data collection and preprocessing are provided in Section 3. The regression target is the average site EUI (AEUI) of residential buildings in Block Groups, which is the most detail census level of the American Community Survey (ACS) in NYC. The results and discussion are shown in Section 4, followed by conclusions in Section 5.

2. Methodology

To address the modeling limitations in linear regression, many non-linear based machine learning algorithms such as Artificial Neural Network (ANN) [21–24] and Support Vector Machines (SVM) [1,25,26] have drawn increasing attention to energy related studies. However, although these two algorithms showed great performance compared to linear regression [1,27], the problem is that they model the regression more like a "black box" [28], and cannot directly interpret the variable importance. To address this issue, this study thereby uses another commonly seen machine learning algorithm, namely Random Forests (RF). It can evaluate the variable importance based on the out-of-bag test of the decision trees in a forest [29]. In addition, RF also reveals good modeling performance in many studies in computer science [29–31].

Several energy related studies have also used RF algorithm in modeling both classification and regression problems. For example, Urraca et al. [32] used RF to model the solar irradiation. Tooke et al. [33] implemented RF to predict the building age and energy consumption. Lahouar and Ben Hadj Slama [34] proposed a RF based model to forecast the day-ahead short term load of electricity. In this study, the RF algorithm is used to model the relationships and investigate the relative importance of the influential features to the AEUI of residential buildings.

2.1. Regression tree

The RF algorithm, designed by Breiman [29], is based on regression trees (or classification trees in other problems). As shown in Fig. 1, the main idea of a regression tree is to split the training data based on the variables like a tree structure. The new data or test case follow the tree structure to the end leaf, and use the average target value of the cases in the end leaf as the regressed value [29]. In the example illustrated in Fig. 1, if there is a test case with an average household size of 2, and average household income of USD 9000, then AEUI is 95 kW h/m^2.

The growing mechanism of a regression tree is a greedy approach [35]. It will search the variables and then split in order to minimize the residual sum of squares (RSS) in Eq. (1).

$$RSS = \sum_{l \in leaves} \sum_{i \in C_l} (y_i - \bar{y}_{C_l})^2 \tag{1}$$

where l is a leave, C_l represents the cases in leaf l, y_i is the observed value and $\bar{y}_{C_l}$ is the average observed value in leaf l. The tree will keep on growing until the stopping criteria is reached. Typically, it is either the number of remaining cases in the end leaf smaller than a threshold or the RSS smaller than a threshold [35].