Developing data-driven models for energy-efficient heating design in office buildings

Zhichao Tian[a], Shen Wei[b], Xing Shi[c,d*]

a. School of Architecture, Southeast University, 210096, Nanjing, P.R. China;

b. The Bartlett School of Construction and Project Management, University London College, WC1E 7HB, London, UK;

c. College of Architecture and Urban Planning, Tongji University, Shanghai, P.R. China;

d. Key Laboratory of Ecology and Energy-saving Study of Dense Habitat (Tongji University), Ministry of Education, P.R. China

**Corresponding author: 20101@tongji.edu.cn*

Abstract

Data-driven methods have been widely applied in the prediction of energy consumption in buildings. However, existing well-established data-driven models can hardly be used for energy-efficient design. This study aims to explore the underlying causes and propose an innovative method to exclusively develop models for energy-efficient design. First, a conventional modeling process was implemented, which includes data precession, statistical analysis, feature selection, and Random Forest classification. Second, an innovative two-step method was proposed to develop data-driven models for energy-efficient design. The first step involves identifying important designable features that can be designed through classification. The second step involves developing classification models for developing energy-efficient design. The experiments were performed on the Commercial Building Energy Consumption Survey (CBECS) dataset that contains 6720 non-residential buildings. The models were built with conventional methods to realize high classification accuracy. However, they cannot be used for energy-efficient design because they lack design variables such as the thickness of wall insulation. The main contributions of this study include the identification of important designable features and development of data-driven models exclusively for energy-efficient design. The proposed method can benefit designers in developing useful data-driven models for building energy-efficient design.

Keywords

Energy-efficient design, data-driven, office buildings, heating energy.

1. Introduction

Buildings utilize approximately 40% of the overall energy consumed in advanced countries [1-3]. In the United Kingdom, space heating accounts for over 60% of total building energy consumption [1]. As a result, designers, energy policymakers, and building owners are aware of the necessity of reducing heating energy by adopting high-performance envelopes, heating, ventilation, and air-conditioning systems(HVAC) and improved operations [4, 5].Furthermore, building energy standards, such as the China Energy Standard for Public Buildings [6], have posed stricter requirements for envelopes and HVAC. Building energy-efficient design is a critical step for realizing low-cost construction and operation [7-9]. Building heating design involves adjusting heating related designable variables including building shape, opaque envelopes, transparent envelops, shading, passive heating, and heating equipment.[10]. In the design stage, design teams should determine the wall's insulation thickness and heating system, which are termed as "designable features". Conversely, some architecture features are pre-fixed, such as building area, functions, and the number of floors.

Several design methods have been proposed in the past several decades to realize energy-efficient design, which mainly refer to heating/cooling load design and simulation-based building energy-efficient design [8, 11]. Energy-efficient design based on heating/cooling load calculation refers to procedures for selecting building variables that minimize the heating load. Given that the heating/cooling load calculation only considers satisfying indoor thermal comfort during winter or summer days, it cannot guarantee high-efficient operation throughout the year. Due to its simplicity, this method is typically used in the early design stage [12]. However, in recent years, designers are abandoning this approach. Simulation-based building energy-efficient design entails detailed dynamic building energy simulation for weighing competing design options. Therefore, the impacts of a variety of energy-efficient measures, such as double-skin facades, can be quantified with energy simulation. However, this method has been widely questioned due to the performance gap, which refers to the huge difference between simulated and measured performances [13-15]. Furthermore, building energy simulation is heavily criticized due to its long modeling time, steep learning curve, and trial-error characteristics [16].

Conversely, data-driven building energy-efficient design (DDBED) has recently attracted significant attention owing to the rapid accumulation of building data. For example, the U.S. Building Performance Database contains over 750,000 entries [17]. Data-driven models can overcome the shortcomings of energy simulation [18]. DDBED generally adopts machine learning methods, typically classified as regression, classification, clustering, and deep learning. In this field, data-driven models have been built in many studies for addressing building energy issues with realistic building data [19-21].

DDBED aims at realizing high-efficient solutions. Hence, in theory, after appropriate training and testing, these data-driven models can be applied to develop high-energy-efficient solutions. For this purpose, the candidate building is assumed to be high-energy-efficient (y=1), and the model is used determine X, as shown in Eq. 1. Classification is suitable for accomplishing this work. If a regression model was built, the design team should evaluate the energy consumption of different design solutions. In the building field, collected data exhibits significant amount of uncertainties due to various reasons such as misunderstanding the meaning of a variable [13, 15]. From the perspective of data-driven energy-efficient design for buildings, a range of values are much more resilient than a value point used by a regression model.

$$f(X) = y \qquad \text{(Eq. 1)}$$

Previously, data-driven models have been built in many studies for building energy prediction [19, 22-25]. However, a large proportion of these models cannot be applied for building design. The main reason may be that those models contain few designable features [19, 23-25]. For example, Robinson et al. [19] deployed 10 regression algorithms in predicting building energy consumption. Even though high predicting accuracies were obtained, those models merely had four features, i.e., building area, heating/cooling degree day, and principle building activity. A building energy simulation model contains a large number of building variables related to architecture, envelopes, HVAC systems, human behavior, and operations [26]. However, data-driven models for energy analysis usually contain several variables [16, 25, 27]. When developing data-driven models for energy prediction, engineers mainly consider prediction accuracy other than energy-efficient design [19]. As a result, these models are good at predicting energy, other than energy-efficient design. Even though several data-driven models utilized designable features for energy analysis, designers lack a specific method for developing data-driven models exclusively for energy-efficient design. Hence, data-driven models have hardly been used for design applications.

In this study, an attempt had been made to accelerate the application of data-driven methods for energy-efficient design. There are three major objectives in this study: 1) identifying determinant features of heating energy consumption for office buildings in the cold region, 2) exploring reasons as why traditionally developed models are hardly applied for building energy-efficient design, 3) proposing an innovative two-step method to develop models for DDBED. The remaining part of this paper consists of five sections. An elaborate literature review on data-driven building energy analysis is given in Section 2. In Section 3, the methodologies of this study, including data preprocessing (Section 3.1), Random Forest (Section 3.2), conventional data-driven model development (Section 3.3), and the proposed two-step method (Section 3.4) are described. Section 4 demonstrates the results of the experiments. Hence, certain in-depth discussions with respect to the results are provided in Section 5. The major conclusions of the study are highlighted in Section 6.

2. Literature Review

Large amounts of measured building energy data can reveal essential information about energy usage patterns [17, 28]. Shahrokni et al. [29] compared the energy-efficient potentials of buildings in different age ranges and concluded that if the existing buildings were retrofitted to satisfy the current codes, the heating energy can be reduced by one-third. Moreover, buildings constructed between 1946 and 1975 were verified to exhibit the largest energy reduction potentials. Household electricity use for heating and cooling was proposed by Wang et al. [30] as a metric to evaluate the effectiveness of China Building Energy Efficiency Standards on residential buildings. The results indicated that households that adopt the energy standards save approximately 41% energy.

Data-driven methods have been applied in several studies to unearth determinant variables of building energy consumption. With energy data of 713 mixed-use buildings in Abu Dhabi, Lin et al. [31] analyzed the impacts of dependent variables on the electricity by using the decision tree algorithm. The results indicated that the chiller quality plays the most significant role in energy consumption. In a study of the energy data of 1052 convenience stores in Taiwan, Kuo et al. [32] integrated data-driven evaluators and optimization search methods to determine the key attributes of energy consumption. They reported that business area lighting, no directing lighting, and the capability of freezer cabinet LED lighting were the top influential factors. In addition to conventional building variables, Ma and Cheng [33] investigated the effects of features related to education, population, economy, environment, and transportation using Random Forest on energy usage data of New York City.

Measured building energy data can also be used to evaluate different design solutions. To date, energy-efficient studies that are conducted with data-driven methods mainly involve energy prediction [21, 23, 33, 34] and energy-saving evaluations for retrofitting [25, 35-37]. A few studies conducted data-driven energy analysis on office buildings. Khayatian et al. [38] proposed building energy retrofit index to support retrofit decision-making. They validated the idea with multi-layer perceptron, autoencoders, and k-means algorithms on 4767 office buildings. Deb et al. [25] built artificial neural networks to predict the pre- and post-retrofit energy savings on 56 office buildings. To overcome the shortcoming of engineers' knowledge and experience, Tian et al. [16] proposed a method to select high-energy-efficient HVAC systems with hundreds of high-energy-efficient buildings via the Bayesian Network algorithm.

Building energy database plays a key role in the DDBED. Currently, the Commercial Building Energy Consumption Survey (CBECS) dataset, the California Commercial End-Use Survey, and Building Performance Database are three well-established building energy datasets in the United States [39]. By using the CBECS dataset, Deng et al. [40] compared the prediction accuracy of several machine learning algorithms, including Support Vector Machine (SVM) and Random Forest, in predicting building end-uses energy. The results indicated that SVM and Random Forest exhibit better results when compared with other statistical and simple machine learning algorithms. To quantify the impact of improved operations, Azar and Menassa [5] conducted a three-phase study, namely data gathering, energy modeling, and parametric variation. In the case study, they applied the proposed method mainly on the CBECS dataset.

3. Methodology

To address the aforementioned problems, this study proposes a two-step approach to develop data-driven models for building energy-efficient design. The first step involves identifying important designable features that can be designed by classification. The second step involves developing classification models for the designable features. Before implementing the proposed method, a conventional classification modeling process is conducted to explore potential reasons as to why existing models are hardly used for energy-efficient design. Classification is an effective technique to predict the energy levels of a building [21, 32]. When compared with regression, classification is more likely to realize high accuracy prediction, as it only predicts several finite categories [32]. Random Forest is adopted to generate data-driven models that can accurately predict the heating energy consumption of the office buildings.

3.1. Data and preprocessing

The analyses were conducted on the CBECS 2012 dataset due to its large sample size (6720 non-residential buildings) and over 100 useful features[41]. This dataset was developed by the U.S. Energy Information Administration with the aim to gain a better understanding of the energy consumption of 560 million existing commercial buildings in the USA. The dataset consists of various building attributes related to the building characteristics and energy consumption. Although it is known as a commercial dataset, the data consists of substantial number of non-commercial buildings, such as hospitals and schools. In this study, only office buildings were included to conduct the experiments because different types of buildings exhibit diverse energy use patterns [23, 42].

For a green building certification, dividing buildings into several categories based on their energy usage intensity (EUI) is a typical practice for calculating their energy scores [32]. In this study, office buildings with heating degree days (based on 65 °F) greater than 2000 were selected to ensure basic heating demands. Based on their heating EUIs, the remaining 814 buildings were classified into low- (75%–100%), medium- (25%–75%), and high-efficient groups (0%–25%).

The main purpose of the models being developed is to design high heating-efficient buildings. Hence, the data labeled as 'high-efficient' and 'low-efficient' were used in training and testing those models. This leads to two evident advantages: 1) reducing the number of categories to two, which can increase prediction accuracy [32]; 2) increasing the difference between the two remaining categories to easily recognize the impact of influential factors. Additionally, the medium buildings are less useful because high energy efficiency is a more important objective in the design state when compared to "medium energy efficiency".

The entries of many features can potentially be missing. Features that miss more than 80% of values were removed from the dataset. In the CBECS dataset, missing value implies that the value is not applicable. In the remaining data, some records that miss important values, such as energy consumption, were also removed. Hence, missing features, mainly related to RENINS, BLDSP, and RENHVC, were filled with 0 (not applicable), which is one of the common practices adopted in machine learning [43]. After preprocessing, 53 features were left. In practice, the pool of candidate features should be further curtailed. Hence, a group of features that may affect the heating energy was selected. Table 1 lists 25 features (in bold) relevant to heating energy-efficient levels (HPLV) of office buildings, with their abbreviations.

Table 1 Building features used in this study

Abbreviation	Explanation	Abbreviation	Explanation
BLDSP	Building shape	**PUBCLIM**	Building America climate region
CENDIV	Census division	**REGION**	Census region
ELHT1	Electricity used for main heating	**RENHVC**	HVAC equipment upgrade
GLSSPC	Percent exterior glass	**RENINS**	Insulation upgrade
HDD65	Heating degree days	**RENWLL**	Exterior wall replacement
HEATPC	Percent heated	**RFCNS**	Roof construction material
MAINHT	Main heating equipment	**SQFT**	Square footage
MONUSE	Months in use	**HTPMPH**	Heat pumps for heating
NFLOOR	Number of floors	**WINTYP**	Window glass type
NWKERC	Number of employees category	**WKHRC**	Weekly hours category
HT2	Energy used for secondary heating	**WLCNS**	Wall construction material
OPNWE	Open on weekend	**YRCON**	Year of construction
BOILER	Building owner	HPLV	Heating energy-efficient levels

3.2. Random Forest classifiers

Building energy consumption is a heterogeneous process, which involves many uncertainties and nonlinear characteristics. Powerful classification algorithms are required to realize high accurate predictions. Fortunately, an increasing number of algorithms are available. Previously, a variety of data-driven algorithms have been applied to analyze building energy, for instance, linear models [22, 31], logistic regression [44], decision trees [21], Artificial Neural Network (ANN) [25, 45], SVM, and ensemble learning [33, 42]. Advanced machine learning algorithms, such as ensemble learning and deep learning, usually outperform simple algorithms such as linear models and decision trees [25, 40, 42, 46].

Ensemble learning adopts multiple simple machine learning models to create a synthesized algorithm that individually outperforms any one of the algorithms that is part of the ensemble, as shown in Fig. 1. Boosting and bagging are two types of commonly used ensemble learning mechanisms. The boosting learning endorses a set of algorithms for converting weaker learners to strong learners based on a proven theory that states that weakly and strongly learnable problems are equal. Bagging deploys multiple bootstrap samples to gain subsets that can be used to train the base learners. Based on the inputs, the ultimate output corresponds to the average output of the base learners [43].

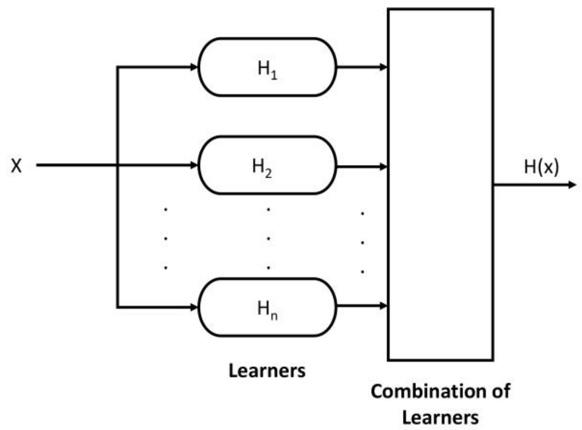

Fig. 1. Mechanism of ensemble learning [47]

As an ensemble learning method, Random Forest uses multiple decision trees as base learners. During training, each tree is distributed with a slice of bootstrap samples. The ultimate predicted result of a test pointer corresponds to the majority voting of the combined classifiers or arithmetic mean of the combined regressors. The randomization in ensemble learning relates to two aspects, i.e., bootstrap sampling and the best split of a node [47]. Furthermore, Sklearn [48], which is a python machine learning algorithm library, was used to build the Random Forest classifiers.

To increase credibility, training and testing sets are typically randomly divided. For example, random 80% of the data are set as training set and the remaining 20% of the data are set as testing set. In this condition, a classifier can accidentally classify the testing set easily. The problem that the testing set contains known training data can lead to overfitting or selection bias [49]. Therefore, several effective solutions, including K-fold cross-validation, bootstrap resampling, and bagging, can be used to solve this problem. K-fold cross-validation divides the original data randomly into k equal-sized parts, which are called folds. In the training stage, each fold is treated as the testing set and remaining k-1 folds are treated as the training set. With this method, a machine learning model is trained and tested K times. The mean value of performance measures (e.g., error rate) and its variance are treated as new performance criteria. Generally, k can be selected as either 5 or 10. Furthermore, K-fold cross-validation is well-received for its effectiveness in minimizing the imperfect effect of partitioning data. In this study, K is set to 4 in the feature selection process. Additionally, in this study, the Area Under Receiver Operating Characteristics Curve (ROC-AUC) is adopted as the classification assessment criterion.

3.3. Conventional model development

The conventional data-driven modeling process entails data preprocessing, statistical analysis, and classification learning. As depicted in the above section, feature selection is an indispensable process for classification modeling of building energy. In machine learning, it involves a process of selecting a subset of relevant features for model development. Filter, wrapper, and embedded methods are three types of feature selection methods [50]. Improving prediction accuracy, producing more cost-effective estimators, and gaining a deeper understanding of the data are the three main objectives of feature selection [50]. This study was planned to determine the extent to which each feature affects the heating energy consumption and to probe the preferable combination of features for predicting heating energy consumption levels. In this study, a conventional classification model development was implemented including statistical analysis and step forward wrapper feature selection.

3.3.1. Statistical analysis

Statistical analysis can be used to not only describe the basic information of the data but also explore the relationship between each feature and energy consumption. In this section, filter methods are adopted to delve into the effect of each feature on heating energy. The filter method calculates the dependence of each feature on the output variable without considering the overall modeling performance [50]. Due to its simplicity, scalability, and empirical success, many studies have adopted this method as a preliminary feature selection method [50]. In this study, the Pearson correlation coefficient and Chi-square testing were adopted for selecting features, by analyzing the relationship between each independent variable, such as HDD65, and the dependent variable, HPLV. The Pearson correlation coefficient was used to quantify the linear correlation between two continuous variables, ranging between -1 and +1, as presented in Eq. (2). The value indicates a positive or negative relationship between variables. As the absolute value increases, the significance of the correlation between the two tested variables increases.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

where cov(X,Y) denotes the covariance of two variables, X and Y; σ_X denotes the standard derivation of X; and σ_Y denotes the standard derivation of Y.

The Chi-square test measures the distribution of a categorical variable in one or more groups. The Chi-square is defined as:

$$\chi^2 = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j} \tag{3}$$

where o_j denotes the observed frequency in event j; e_j denotes the expected frequency in event j; K denotes the total number of events. The sample distribution of χ^2 is close to a Chi-square distribution. The p-value of the Chi-square determines whether to accept the null hypothesis. In this study, the Chi-square test is conducted for each feature on high- and low-energy-efficient buildings.

3.3.2. Step forward wrapper

Although features related to heating energy consumption can be selected with filter methods, a major limitation of filter methods is that they ignore the overall performance of the developed Random Forest models. To tackle this issue, step forward feature selections were deployed. For this method, the first step involves evaluating the classification performance for every feature. Specifically, the feature that provides the best performance is appointed as the first feature. In the second step, each of the remaining features are grouped sequentially with the first feature to determine the best combination of two features. These types of trials involving the aforementioned combinations are repeated many times until all features are ranked. In this study, the step forward wrapper method was carried out with Mlxtend [51], a python library for data analysis and machine learning.

3.4. Proposed model development

The conventional model development emphasizes on the prediction accuracy other than the practicability of energy-efficient design. Previous studies showed that supervised learning models can be built only with several features [25]. This implies that if a feature was not used by a model, then it cannot be designed. Hence, it is necessary to identify features that can be designed via classification and to develop classification models for designing the features. In this section, a two-step procedure was proposed to fulfill this plan. The first step involves excavating features, termed as important designable features, that can be designed by classification. The second step involves developing classification models that mainly adopt important designable features. The following two sub-sections discuss and describe detailed approaches to fulfill the aforementioned steps.

3.4.1. Identifying important designable features with SHAP values

Before applying data-driven models for building energy-efficient design, it is important to identify as to which features can be designed. We assumed that features that significantly impact the outcome of Random Forest models are important features. Important designable features correspond to a union of important features and designable features, as shown in Fig. 2. Hence, weighting the effect of each feature on the outcome is prioritized for identifying important designable features.

Unlike decision tree and linear regression, the outcomes of advanced machine learning models, including Random Forest, are hard to interpret [52]. To solve this problem, Lundberg et al. proposed the SHapley Additive exPlanations (SHAP) method to explain the outcomes of advanced machine learning models [53]. This method allows engineers to quantify the impact of each feature on outcomes of a model. In this study, this method was used to quantify the impacts of each feature on energy prediction. Then, important designable features were selected based on their SHAP values.

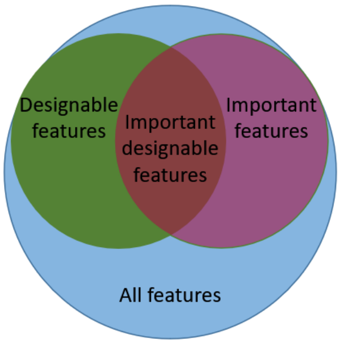

Fig. Relationship between designable features and important features

3.4.2. Models for energy-efficient design

Once the important designable features are identified, the next step involves developing a Random Forest model for designing them. When considering existing feature selection approaches, the optimization-based wrapper feature selection is a practical choice because it can explore the effects of different combinations of features. Furthermore, it can set important designable features as default. Given that the prediction accuracy is the sole objective of optimization, Single Objective Genetic Algorithm (SOGA) can be used to explore feasible feature combinations. Inspired by Darwin's evolution theory, Genetic Algorithm was introduced for generate high-quality solutions with operators such as reproduction, mutation, recombination, and selection. GA has been successfully applied to solve feature selection problems [54-56]. In this study, the optimization process aims at minimizing the objective function by attempting various input values wherein the statue of each feature is represented with a value of either 0 or 1, as per the binary system. Typically, the initial population is randomly seeded [57]. All variables are initialized to 1. Table 2 lists the detail settings of SOGA. In the proceeding population, a fitness function is used to generate a new generation based on a part of existing generations. A merit function penalizes unfeasible variables by using an exterior function. Replacement sets the mechanism for replacing certain selected members to continue the next generation. A favorable feasible type replacement firstly considers feasible as a selection standard. If it cannot realize a winning solution, then it considers the fitness value. This implies that a favor feasible type replacement enforces the fitness assessor. The crossover type defines as to how the genetic information of two parents is used for generating a child. Shuffle random type crossover randomly selects a design variable from two or more parents. Each variable is expected to be equally distributed between 0 and 1. An offset_uniform mutation type enables the mutation of a variable value by using a uniform distribution. Furthermore, Dakota toolkit, which is developed by Sandia National Laboratory of U.S., provides a variety of iterative methods and meta-algorithms for optimization, sensitive analysis, uncertainty analysis, and parameter studies [58]. Due to its open source characteristics and ready to use python API, Dakota optimization engine was deployed to fulfill the SOGA process.

Table 2 Detailed settings of SOGA

Fitness type	Replacement type	Convergence type	Crossover type	Mutation type
Merit function	Favor feasible	Best fitness tracker	Shuffle random	Offset uniform

4. Results

4.1. Traditional modeling

As previously stated, the Pearson correlation coefficient is used to measure the correlation between dependent and independent continuous variables. The significance of the correlation is proportional to its absolute value. Before conducting the correlation analysis, logarithmic transforms were implemented on features that are akin to exponential distribution, including SQFT, NWKERC, and MONUSE. Then, these continuous features were implemented with normality tests to verify whether they follow the normal distribution. The results indicated that all continuous features passed the normality test with the exception of GLSSPC. After these analyses, the Pearson correlation coefficient analysis was conducted on continuous variables and heating energy consumption intensity (HEUI). Table 3 lists the Pearson correlation coefficients for those continuous variables. If the threshold of the Pearson correlation coefficient is set as 0.10 to decide the significance of contribution as suggested in [59], then HDD65, YRCON, and HEATP are considered as important features of the building heating energy consumption.

Table 3 Pearson correlation coefficients between the continuous variables and HEUI

Feature	SQFT	**HDD65**	NWKERC	NFLOOR	WKHRSC	**YRCON**	MONUSE	**HEATP**
Corr	-0.12	0.25	-9.8e-2	-7.2e-2	5.2e-3	-0.19	5.9e-2	0.10

Furthermore, the Pearson correlation coefficients of pair-wise features were also calculated. Table 4 lists the pair-wise features, whose correlation coefficients are higher than 0.5 and present explanations for a significant correlation between these features.

Table 4 Pair-wise features (Corr>0.5)

Feature	Feature	Explanation
SQFTC	NWKERC	As the building size increases, the number of people who may work in this building increases.
REGION	CENDIV	Both used to describe buildings' locations.
RENWLL	RENHVC	Once a building was renovated, the HVAC system and exterior wall could be retrofitted.
RENWLL	RENINS	Renovations of insulation is also a type of renewing the wall.
RENHVC	RENINS	Once a building was renovated, the HVAC system and the exterior wall could be retrofitted.

The p-value of the Chi-square test provides evidence of whether the tested feature is statistically significant to HPLV. Table 5 lists p-values for each categorical feature. If the significance level is set as 0.01, then the selected features correspond to BOILER, CENDIV, ELHT1, MAINHT, HT2, PUBCLIM, MAINCL, REGION, RENHVC, and HTPMPH.

Table 5 P-values for Chi-test for categorical features in different heating energy-efficient groups

Feature	BLDSP	CENDIV	ELHT1	MAINHT	HT2	OPNWE	BOILER	PUBCLIM
P-value	8.87e-2	8.94e-10	1.14e-21	2.31e-8	7.99e-6	0.107	8.73e-5	1.74e-13
Feature	REGION	RENHVC	RENINS	RENWLL	RFCNS	HTPMPH	WINTYP	WLCNS
P-value	2.26e-8	8.86e-3	1.41e-2	1.19e-2	0.105	2.20e-8	3.90e-2	0.226

Table 6 lists the selected features at each step and their corresponding prediction accuracy. It can be observed that the overall modeling accuracy varies at each step, and it realizes the highest accuracy at Step 2. Hence, HPLV can be predicted only with two features, i.e., ELHT1 and PUBCLIM.

Table 6 Features raised by the step forward wrapper method

Step	1	2	3	4	5	6	7
Accuracy	0.770	0.844	0.838	0.844	0.843	0.845	0.855
Feature	ELHT1	PUBCLIM	HT2	HTPMPH	CENDIV	MAINHT	HEATP
Step	8	9	10	11	12	13	…
Accuracy	0.847	0.840	0.839	0.832	0.818	0.804	…
Feature	REGION	BOILER	WINTYP	YRCON	HDD65	RENHVC	…

Table 6 shows that the Random Forest model can be built only with two features, i.e., ELHT1 and PUBCLIM. It is likely that ELHT1 undermines the impact of MAINHT because ELHT1 is a derivation of MAINHT, and thereby represents whether a building uses electricity for heating. For this reason, ELHT1 was deleted from candidate features, and the model was thus rebuilt. Table 7 demonstrates the features raised by the step forward wrapper feature selection method after deleting ELHT1. Table 7 shows that the best model is realized with PUBCLIM, MAINHT, and HTPMPH.

Table 7 Features selected by the step forward wrapper method without ELHT1

Step	1	2	3	4	5	6	7
Accuracy	0.731	0.815	0.821	0.821	0.806	0.809	0.809
Feature	PUBCLIM	MAINHT	HTPMPH	BOILER	HT2	CENDIV	REGION
Step	8	9	10	11	12	13	…
Accuracy	0.802	0.799	0.815	0.796	0.789	0.789	…

Feature	HEATP	HDD65	YRCON	RENHVC	MAINCL	CENDIV	…

4.2. Two-step modeling

Figure 3 shows the distribution of SHAP values of each feature for each data point. In this figure, features were ranked based on the summations of their SHAP values, and only the top 20 features were plotted. Figure 4 shows the mean value of SHAP values for the top 20 features. This diagram clearly demonstrates that ELHT1 exhibits the highest impact on the Random Forest classification. Within the top 10 features, only MAINHT and BOILER are designable. Hence, the developed classification models can mainly be used to design these two features. Fig. 5, the intersection of two sets of feature groups demonstrates these important designable features. Given that the BOILER is derived from the MAINHT, Random Forest models can be developed just for designing MAINHT in the next step.

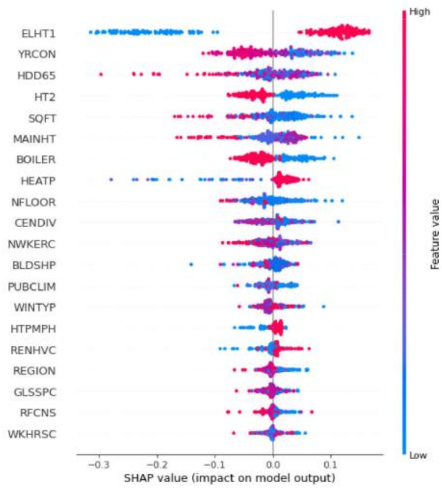

Fig. 3. Scatter diagram of the SHAP values of each feature

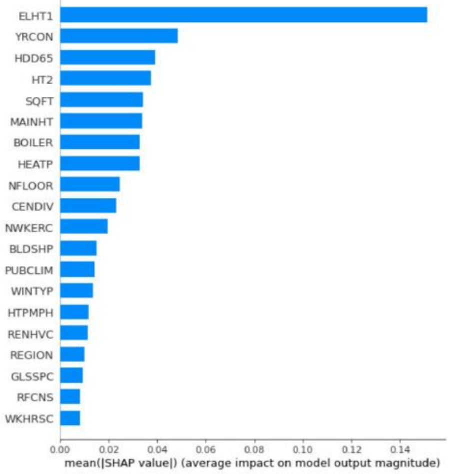

Fig. 4 Bar diagram showing the mean of SHAP values for different features

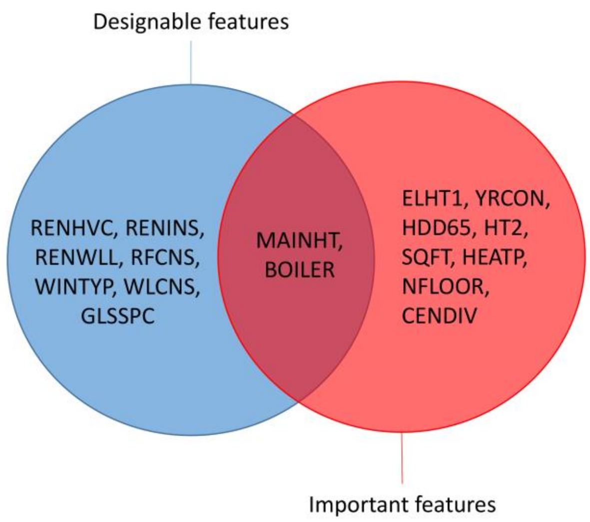

Fig. 5 Intersection of two sets of feature groups

Table 8 shows the feature combinations selected by the SOGA-based wrapper method. The results indicated that these models exhibit acceptable accuracy. Hence, they can be applied to design MAINHT, i.e., the main heating equipment.

Table 8 Top 4 Random Forest models developed by the SOGA-based wrapper feature selection

Model ID	Feature combination	ROC_AUC
303	MAINHT, CENDIV, PUBCLIM, WLCNS, HEATP, NFLOOR	0.793
301	MAINHT, CENDIV, PUBCLIM, WLCNS, NFLOOR	0.786
761	MAINHT, PUBCLIM, REGION, WINTYP, WLCNS, MONUSE, HDD65, NFLOOR	0.785
566	MAINHT, CENDIV, WINTYP, HEATP, NFLOOR	0.782

5. Discussion

In the conventional modeling process, statistical analysis was used to sort determinant features related to heating energy consumption. The Pearson correlation coefficient method was used to identify three important features, i.e., HDD65, YRCON, and HEATP. The Chi-square test targeted CENDIV, PUBCLIM, MAINCL, MAINHT, REGION, HT2, ELHT1, BOILER, and HTPMPH, which exhibit a strong relationship with heat energy consumption. One of the drawbacks of these methods is that they failed to identify the effect of specific observations of the determinant features. For example, it is not clear as to which heating system is most frequently used in low-energy-efficient buildings. A visual solution involves comparing the distribution of observations in different groups of buildings. Figure 6 describes the distribution of each heating system in high-, medium-, and low-energy-efficient buildings. Hence, it can be observed that many more heat pumps were installed in the high-energy-efficient buildings.

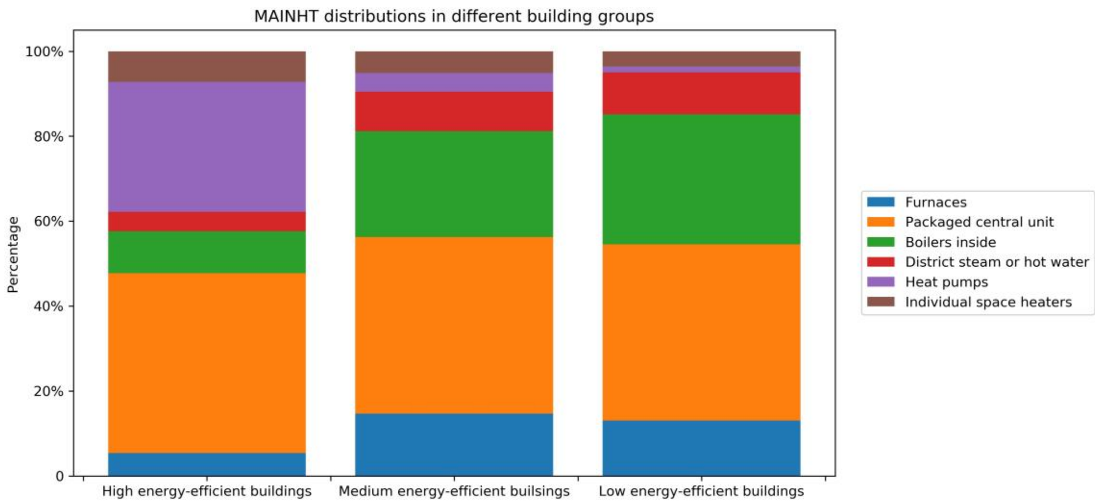

Fig. 6 Distribution of heating systems in different building groups

The best model generated by the step forward method exhibits only two features, i.e., ELHT1 and PUBCLIM. Although the ROC_AUC value of 0.844 is high enough for predicting whether a building is high-efficient or low-efficient, it is almost impossible to conduct energy-efficient design due to the lack of designable features. After deleting ELHT1 from the candidate features, MAINHT was selected by the step forward wrapper method as a key feature in predicting HPLV. This process requires a good understanding of the meaning of every feature. In practice, it is a trial-and-error process that can be significantly time-consuming and unconvincing.

The SHAP method can eliminate the phenomenon that the features selected by the wrapper feature selection method undermine the impacts of other features. In this study, the ELHT1 undermines the effect of MAINHT and BOILER. Based on the Chi-square test, designers are required to provide a significance level for the SHAP values to determine the important features. Although, these Random Forest models contain other designable features, Random Forest models should not be used to design the important features.

6. Conclusions

In this study, the development of data-driven models for building energy-efficient design is explored. The traditional data-driven modeling process successfully led to several Random Forest models that realize high prediction accuracy. However, these models cannot be used for building energy-efficient design because they lack designable features. The proposed two-step modeling method can be used to identify important designable features and develop Random Forest models for designing them. Based on the ROC_AUC values, the Random Forest models exhibited acceptable results. The results indicated that Random Forest models can be used to design the main heating equipment (MAINHT), a dominant feature of heating energy consumption in office buildings in the cold region.

The proposed techniques are useful for policymakers and building energy consultants. It can aid the local governments to formulate energy policies for important designable features. Furthermore, the techniques allow designers to build classification models for building energy-efficient design for applications other than just energy prediction.

Given that only the design of important designable features was addressed in the present study, it should be examined to develop suitable methods for building energy-efficient design for other designable features. Possible avenues for pursuing this include recommending design solutions for non-determinant features with unsupervised learning.

Acknowledgment

This paper is financially supported by the Ministry of Science and Technology of China (Project number: 2016YFC0700102), Scientific Research Foundation of Graduate School of Southeast University (YBJJ1801).

Reference

1. Perez-Lombard, L., J. Ortiz, and C. Pout, *A review on buildings energy consumption information.* Energy and Buildings, 2008. **40**(3): p. 394-398.
2. Recast, E.P.B.D., *Directive 2010/31/EU of the European Parliament and of the Council of 19 May, 2010 on the energy performance of buildings.* Official Journal of the European Union, 2010. **153**: p. 13-35.
3. Sieminski, A.J.E.I.A., *International energy outlook.* 2014. **18**.
4. Fasano, G., M.J.B. Zinzi, and Environment, *Optimisation of opaque components of the building envelope. Energy, economic and environmental issues.* 2006. **41**(8): p. 1001-1013.
5. Azar, E. and C.C.J.E.P. Menassa, *A comprehensive framework to quantify energy savings potential from improved operations of commercial building stocks.* 2014. **67**: p. 459-472.
6. Ministry of Housing and Urban-Rural Construction of the People's Republic of China, *Design standard for energy efficiency of public buildings*. 2015, China Architecture& Building Press. p. 2.
7. Omer, A.M.J.R. and s.e. reviews, *Renewable building energy systems and passive human comfort solutions.* 2008. **12**(6): p. 1562-1587.
8. Shi, X., *Design optimization of insulation usage and space conditioning load using energy simulation and genetic algorithm.* Energy, 2011. **36**(3): p. 1659-1667.
9. Negendahl, K., T.R.J.E. Nielsen, and Buildings, *Building energy optimization in the early design stages: A simplified method.* 2015. **105**: p. 88-99.
10. Pacheco, R., et al., *Energy efficient design of building: A review.* 2012. **16**(6): p. 3559-3573.
11. Hien, W.N., L.K. Poh, and H. Feriadi, *The use of performance-based simulation tools for building design and evaluation - a Singapore perspective.* Building and Environment, 2000. **35**(8): p. 709-736.
12. Carlos, J.S., M.C.J.E. Nepomuceno, and Buildings, *A simple methodology to predict heating load at*

an early design stage of dwellings. 2012. **55**: p. 198-207.
13. van den Brom, P., A. Meijer, and H. Visscher, *Performance gaps in energy consumption: household groups and building characteristics.* Building Research & Information, 2018. **46**(1): p. 54-70.
14. de Wilde, P., *The gap between predicted and measured energy performance of buildings: A framework for investigation.* Automation in Construction, 2014. **41**: p. 40-49.
15. Zou, P.X.W., et al., *Review of 10 years research on building energy performance gap: Life-cycle and stakeholder perspectives.* Energy and Buildings, 2018. **178**: p. 165-181.
16. Tian, Z., et al., *An application of Bayesian Network approach for selecting energy efficient HVAC systems.* Journal of Building Engineering, 2019: p. 100796.
17. Mathew, P.A., et al., *Big-data for building energy performance: Lessons from assembling a very large national database of building energy use.* Applied Energy, 2015. **140**: p. 85-93.
18. Deb, C. and S.E. Lee, *Determining key variables influencing energy consumption in office buildings through cluster analysis of pre-and post-retrofit building data.* Energy and Buildings, 2018. **159**: p. 228-245.
19. Robinson, C., et al., *Machine learning approaches for estimating commercial building energy consumption.* Applied Energy, 2017. **208**: p. 889-904.
20. Aksoezen, M., et al., *Building age as an indicator for energy consumption.* Energy and Buildings, 2015. **87**: p. 74-86.
21. Yu, Z., et al., *A decision tree method for building energy demand modeling.* Energy and Buildings, 2010. **42**(10): p. 1637-1646.
22. Wong, I.L., et al., *Classification and energy analysis of bank building stock: A case study in Curitiba, Brazil.* Journal of Building Engineering, 2019. **23**: p. 259-269.
23. Wang, J.C., *A study on the energy performance of school buildings in Taiwan.* Energy and Buildings, 2016. **133**: p. 810-822.
24. Huebner, G., et al., *Understanding electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes.* Applied Energy, 2016. **177**: p. 692-702.
25. Deb, C., S.E. Lee, and M. Santamouris, *Using artificial neural networks to assess HVAC related energy saving in retrofitted office buildings.* Solar Energy, 2018. **163**: p. 32-44.
26. Hensen, J.L. and R. Lamberts, *Building performance simulation for design and operation*. 2012: Routledge.
27. Amasyali, K. and N.M. El-Gohary, *A review of data-driven building energy consumption prediction studies.* Renewable & Sustainable Energy Reviews, 2018. **81**: p. 1192-1205.
28. Hsu, D., *How much information disclosure of building energy performance is necessary?* Energy Policy, 2014. **64**: p. 263-272.
29. Shahrokni, H., F. Levihn, and N. Brandt, *Big meter data analysis of the energy efficiency potential in Stockholm's building stock.* Energy and Buildings, 2014. **78**: p. 153-164.
30. Wang, X., et al., *Do residential building energy efficiency standards reduce energy consumption in China?–A data-driven method to validate the actual performance of building energy efficiency standards.* 2019. **131**: p. 82-98.
31. Lin, M., A. Afshari, and E. Azar, *A data-driven analysis of building energy use with emphasis on operation and maintenance: A case study from the UAE.* Journal of Cleaner Production, 2018. **192**: p. 169-178.
32. Kuo, C.F.J., C.H. Lin, and M.H. Lee, *Analyze the the energy consumption characteristics and affecting*

factors of Taiwan's convenience stores-using the big data mining approach. Energy and Buildings, 2018. **168**: p. 120-136.

33. Ma, J. and J.C.P. Cheng, *Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests.* Applied Energy, 2016. **183**: p. 193-201.
34. Li, Q., P. Ren, and Q. Meng. *Prediction model of annual energy consumption of residential buildings*. in *2010 international conference on advances in energy engineering*. 2010. IEEE.
35. Marasco, D.E. and C.E. Kontokosta, *Applications of machine learning methods to identifying and predicting building retrofit opportunities.* Energy and Buildings, 2016. **128**: p. 431-441.
36. Hamilton, I.G., et al., *Energy efficiency uptake and energy savings in English houses: A cohort study.* Energy and Buildings, 2016. **118**: p. 259-276.
37. Walter, T. and M.D. Sohn, *A regression-based approach to estimating retrofit savings using the Building Performance Database.* Applied Energy, 2016. **179**: p. 996-1005.
38. Khayatian, F., L. Sarto, and G. Dall'O, *Building energy retrofit index for policy making and decision support at regional and national scales.* Applied Energy, 2017. **206**: p. 1062-1075.
39. Ye, Y., et al., *A comprehensive review of energy-related data for US commercial buildings.* 2019.
40. Deng, H.F., D. Fannon, and M.J. Eckelman, *Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata.* Energy and Buildings, 2018. **163**: p. 34-43.
41. EIA. *COMMERCIAL BUILDINGS ENERGY CONSUMPTION SURVEY (CBECS)*. 2019 [cited 2019 21-Apr.]; Available from: https://www.eia.gov/consumption/commercial/data/2012/index.php?view=microdata.
42. Hsu, D., *Identifying key variables and interactions in statistical models of building energy consumption using regularization.* Energy, 2015. **83**: p. 144-155.
43. Han, J., J. Pei, and M. Kamber, *Data mining: concepts and techniques*. 2011: Elsevier.
44. Kontokosta, C.E., *Modeling the energy retrofit decision in commercial office buildings.* Energy and Buildings, 2016. **131**: p. 1-20.
45. Jing, R., et al., *A study on energy performance of 30 commercial office buildings in Hong Kong.* Energy and Buildings, 2017. **144**: p. 117-128.
46. Papadopoulos, S. and C.E. Kontokosta, *Grading buildings on energy performance using city benchmarking data.* Applied Energy, 2019. **233**: p. 244-253.
47. Zhang, C. and Y. Ma, *Ensemble machine learning: methods and applications*. 2012: Springer.
48. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* Journal of machine learning research, 2011. **12**(Oct): p. 2825-2830.
49. Cawley, G.C. and N.L.C. Talbot, *On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation.* Journal of Machine Learning Research, 2010. **11**: p. 2079-2107.
50. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* Journal of machine learning research, 2003. **3**(Mar): p. 1157-1182.
51. Raschka, S. *MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack*. 2019 [cited 2019 Jan-1]; Available from: http://rasbt.github.io/mlxtend/.
52. Štrumbelj, E., I.J.K. Kononenko, and i. systems, *Explaining prediction models and individual predictions with feature contributions.* 2014. **41**(3): p. 647-665.
53. Lundberg, S.M. and S.-I. Lee. *A unified approach to interpreting model predictions*. in *Advances in neural information processing systems*. 2017.

54. Shah, S.C. and A. Kusiak, *Data mining and genetic algorithm based gene/SNP selection.* Artificial Intelligence in Medicine, 2004. **31**(3): p. 183-196.
55. Jirapech-Umpai, T. and S. Aitken, *Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes.* Bmc Bioinformatics, 2005. **6**.
56. Xuan, P., et al., *Genetic algorithm-based efficient feature selection for classification of pre-miRNAs.* Genetics and Molecular Research, 2011. **10**(2): p. 588-603.
57. Whitley, D., *A genetic algorithm tutorial.* Statistics and computing, 1994. **4**(2): p. 65-85.
58. Adams, B.M., et al., *DAKOTA, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: version 5.0 user's manual.* Sandia National Laboratories, Tech. Rep. SAND2010-2183, 2009.
59. Spiegel, M.R. and L.J. Stephens, *Statistics*. 2018: McGraw-Hill New York.