Applied Energy 183 (2016) 182–192

Contents lists available at ScienceDirect

Applied Energy

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

Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology

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

- A GIS integrated data mining framework was proposed for estimating building EUI.
- A case study on 3640 residential buildings in NYC was used to test the framework.
- The proposed framework can help produce lower estimation error than past studies.
- A comparative study on the feature selection and regression methods was conducted.
- The model built by SVM on the features selected by Elastic Net has the lowest MSE.

ARTICLE INFO

Article history:
Received 28 January 2016
Received in revised form 25 July 2016
Accepted 14 August 2016

Keywords:
Artificial Neural Network (ANN)
Big Data
Energy use intensity (EUI)
Feature selection
Geographic information system (GIS)
Support Vector Regression (SVR)

ABSTRACT

Buildings are the major source of energy consumption in urban areas. Accurate modeling and forecasting of the building energy use intensity (EUI) in the urban scale have many important applications, such as energy benchmarking and urban energy infrastructure planning. The use of Big Data technology is expected to have the capability of integrating a large number of predictors and giving an accurate prediction of the energy use intensity of buildings in the urban scale. However, past research has often used Big Data technology in estimating energy consumption of a single building rather than the urban scale, due to several challenges such as data collection and feature engineering. This paper therefore proposes a geographic information system integrated data mining methodology framework for estimating the building EUI in the urban scale, including preprocessing, feature selection, and algorithm optimization. Based on 216 prepared features, a case study on estimating the site EUI of 3640 multi-family residential buildings in New York City, was tested and validated using the proposed methodology framework. A comparative study on the feature selection strategies and the commonly used regression algorithms was also included in the case study. The results show that the framework was able to help produce lower estimation errors than previous research, and the model built by the Support Vector Regression algorithm on the features selected by Elastic Net has the least cross-validation mean squared error.

1. Introduction

The global contribution from buildings in regard to energy consumption has steadily increased to 20%–40% in developed countries, and has exceeded the other major sectors including industrial and transportation [1]. Especially in some dense urban areas like New York City, buildings account for a staggering 94% of electricity usage and 75% of greenhouse gas (GHG) emissions [2]. Therefore, studying the energy consumption of buildings is crucial in order to better understand the patterns and characteristics, and thus help reduce the overall energy consumption in the urban scale. Accurate modeling and forecasting of building energy consumption enables better energy management and efficient applications, such as the propagation of early stage design decisions [3], the estimation of improvements to building energy performance [4], the optimization of building HVAC systems [5], and the urban energy infrastructure planning [6].

Previous research on the energy consumption of buildings in the urban scale usually focused on the influence of a single kind of feature on the energy consumption. Examples include climate [7–9], urban form [10–12], residential density [13–16], income of the residents [17,18], etc. However, there is a lack of research that integrates all the possibly related features, compares the feature influence and estimates the urban scale energy consumption. To narrow this gap, a large variety of related data must be collected

* 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.2016.08.079

and analyzed, and such kind of research is often referred to as the so-called “Big Data” analytics [19].

“Big Data” analytics or data mining often refers to the implementation of machine learning algorithms, statistical methods, and artificial intelligence technologies to discover hidden knowledge from large datasets [20]. Thanks to the growing availability of various data sources, increasing amount of studies in the energy related research fields have used data mining technology to solve research problems. For example, Koo and Hong [21] used decision tree and other data mining techniques to develop a dynamic operational rating system in energy performance certificates for existing buildings. Siami-Irdemoosa and Dindarloo [22] implemented neural networks to predict the fuel consumption of mining dump trucks. Jain et al. [2] used Support Vector Regression to forecast the energy consumption of multi-family residential buildings. Long et al. [23] used four commonly seen regression models to estimate the daily solar power. However, most of the existing studies focus on the building scale energy consumption, and limited research has been conducted to estimate the building energy consumption in the urban scale. Howard et al. [6] estimated the urban scale building energy consumption by end use. However, some important factors, such as building age, were not considered in their study. Hsu [19] modeled the energy use intensity of buildings in the city scale, but his work focused more on identifying important variables in statistical models.

In addition, there are usually three major challenges for data mining tasks in the urban scale [19,20]. First is the pre feature engineering, including data collection, integration and transformation. This stage is always a time consuming part, because the study may include thousands of observations and hundreds of features/variables. In most cases, those datasets are collected from different sources with different indexes, and may not be easy to join [20]. In fact, some studies extracted the key data from commercial websites and had to develop additional programs to add and join the features [24,25]. For energy estimation problems in the urban scale, it is very likely that the involved datasets or features are indexed by geo ids, or presented in a shapefile. These kinds of datasets can be easily joined using geographic information systems (GIS). A GIS is a system designed to capture, store, manipulate, analyze, manage, and present all types of spatial or geographical data [26]. Many existing energy related literatures have implemented GIS in their studies. However, the majority used it as a visualization or problem engineering tool [27–29], and there is a lack of studies that utilized its powerful capability in managing geographical data in Big Data analysis in the energy field. Hsu [19] used GIS when joining his data, but he did not further explore its functionality in filling missing values and feature generation.

The second challenge of data mining in the urban scale is the feature selection. Among the collected and preprocessed features, not all of them are good for the regression model. In fact, there are always a number of redundant or irrelevant features, which may add noise to the model and thus reduce the performance. As a result, proper selection of the key variables is usually one of the most important steps in data mining studies [19,20,24,30]. The third challenge is the selection of regression/classification algorithms. There are many regression algorithms that have been implemented by previous research in energy related studies, such as Artificial Neural Networks (ANN) [22,25,30], Support Vector Regression (SVR) [2,31,32], and Generalized Linear Models (GLM) [19,33]. The optimal choice of the regression algorithm for different problems is very likely to be different. As a result, a comparative study of those algorithms on data mining tasks in the urban scale becomes an important step.

In summary, to address the gaps and challenges mentioned above, this paper proposes a methodology framework to estimate building energy consumption in the urban scale by implementing GIS and Big Data technology. The objectives are not only to explore an effective and efficient framework to study urban scale energy related problems based on Big Data, but also to conduct a comparative analysis on the commonly seen Big Data technology, including feature selection techniques and regression algorithms. Details of the proposed framework are introduced in Section 2. The case study used to test the effectiveness of the proposed framework is presented in Section 3. Discussion and conclusions follow.

2. Methodology framework

As is shown in Fig. 1, the proposed framework consists of five major parts: (1) data collection, (2) preprocessing, (3) feature selection, (4) validation and parameter identification, and (5) regression. The data collection and preprocessing parts are integrated with GIS. The geocode indexed datasets in GIS can make processes like feature integration and feature generation much easier. The feature selection process and regression models are integrated with the cross validation and the grid search to produce more objective and stable outcomes. Details of the methodology and algorithms in the proposed framework are introduced below.

2.1. Data collection and preprocessing based on GIS

The proposed framework, as shown in Fig. 1, integrates GIS as a hub of data collection and preprocessing. The benefits of this can be summarized into three aspects. First is the data integration. GIS can join unrelated information or datasets by using location as the key index variable. For example, the demographical and housing datasets in the American Community Survey 5-year estimates can be easily joined to individual buildings in a city using shapefiles and geocoding tools in GIS. Traditional ways may require writing a program using Java or Python to geocode the street address through the Google application program interface [24].

Second is the missing values. There will always be features with missing values in Big Data related studies. Commonly seen methods use mean, median or the mode number to fill the blanks [20]. GIS offers another possible approach, which is to use the prediction value from multiple linear regression based on geocodes (latitude and longitude), and this way of filling missing values is more reasonable to the majority of the geo indexed features. Examples are introduced in Section 3.1.

Last is the feature generation. In many cases, the collected raw data may not directly contain the information required. Therefore, further transformation or feature generation is required [20]. For example, urban vegetation may affect the heat island effect and thus influence the building energy consumption, and this is one of the main purposes of having the central park in Manhattan, New York City. Instead of the vegetation index of the place where the building is exactly located, the average vegetation index of the region where the building is located may have more impact on the building energy use intensity. With the help of GIS and shapefiles (map files in GIS), these kinds of features can be easily generated and joined to the dataset. Details and examples are also covered in Section 3.1.

2.2. Feature selection

Urban scale energy estimation using Big Data analytics may contain hundreds of attributes, many of which may be irrelevant or redundant [19,20,24]. In fact, those features may not only slow down the calculation process but also mix the weight of important features, and thus lower the model performance. As a result, a proper feature selection process is a must. Commonly used feature

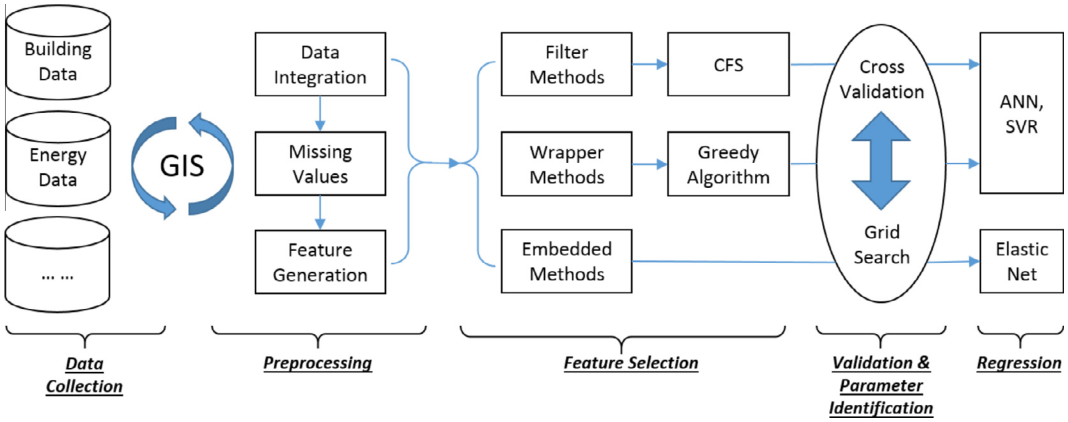

Fig. 1. The proposed methodology framework.

selection approaches in computer science can be divided into three aspects, namely (1) filter methods, (2) wrapper methods, and (3) embedded methods [34]. This study will compared the performance of the representative feature selection approaches in these three aspects. Details are as follows.

2.2.1. Filter methods

Filter methods basically rank the features based on their relationship with the target variable. Different methods just vary in the way of measuring the "relationship". Traditional filter methods are univariate, and only consider the "distance" between one feature to the target, using measurements such as information gain and correlation coefficients [34]. Compared to the wrapper methods, they are computationally efficient but it is easy to get poor subsets because the methods ignore the relationship between features. To address this issue, this study uses the well-known Correlation Feature Selection (CFS) as the filter method due to its high computational efficiency and better performance over univariate methods [24,30,35].

CFS filters features are based on the concept that good subsets contain features highly correlated with the target, yet uncorrelated to each other. The criteria is given in Eq. (1) [35].

$$S = \underset{S}{\operatorname{argmax}}\left[\frac{\sum_{i=1}^{k} Cor_{ti}}{\sqrt{k + 2\sum_{i\neq j}^{k} Cor_{ij}}}\right] \qquad (1)$$

where S is the set of selected features with size k, Cor_{ij} is the correlation between feature i and feature j, and is given by the Pearson's correlation coefficient in this study. t refers to the target. By setting membership indicator functions to the features, Eq. (1) can be transformed into a combinatorial problem and can be solved using branch-and-bound algorithms [35].

2.2.2. Wrapper methods

The wrapper methodology is a simple but powerful way to select features. Unlike the filter methods, the basic idea is to test the model performance of the subset candidates, and select the best one. Since every candidate will run through the model, the drawback of this method is the computation time [34]. In practice, to conduct a wrapper method, one needs to define (1) which model to use and how to assess the model performance, and (2) how to search the space of all possible variable subsets. The first point is simply based on the problem being studied and the algorithm being tested, while in regard to the second point, several search strategies can be used, such as the genetic algorithm, best-first search, and the greedy algorithm [34]. Among these search algorithms, the greedy algorithm is particularly computationally advantageous and robust against overfitting [34], and it also comes close to the optimal solution [20]. As a result, the greedy algorithm is selected as the search algorithm in the wrapper method in this study.

There are basically two types of greedy algorithms in feature selection: greedy forward selection and greedy backward elimination [34]. The greedy forward selection starts with an empty "selected" set, and the best of the original features are determined by the model and added to the selected set. At each subsequent iteration, the best of the remaining original features is added to the set [20]. The greedy backward elimination is just the opposite. It starts with all the original features as the "selected" set, and at each iteration, it removes the worst feature remaining in the set [20]. In this study, since the key features in estimating the building energy use intensity are a small portion of all the prepared features, starting from an empty set to select key features would be more computationally advantageous, and therefore greedy forward selection is used as the wrapper method in the case study in Section 3.

2.2.3. Embedded methods

Embedded methods refer to the feature selection procedure that has already been designed into the regression/classification algorithm itself. For example, the Decision Tree algorithm uses information theory to split and grow the decision tree based on the selected features [36]. In many cases, the pruned Decision Tree will only involve a subset of features, and thus the feature selection is implicitly built into the algorithm. However, this algorithm usually works better for learning discrete valued functions [36]. Another example is the Elastic Net algorithm, which is also known as the LASSO method when the parameter α equates 1. Because of the fast calculation and good regression performance, this method is widely used in different research fields such as computer science [37], ecology [38], and energy [19,33]. Details of the algorithm is covered in Section 2.4.1.

2.3. Validation and parameter identification

Several past studies on estimating the energy consumption directly assess their prediction performance on the training data, or just the test data [6,39]. This way of assessment has a high risk of overfitting [20]. A more objective way of assessing the model performance is cross-validation [20,25]. As shown in Fig. 2, a 10-fold cross validation means the whole dataset is divided into 10 mutually exclusive subsets (or folds). The tested regression algorithm will run 10 rounds on the dataset. In each run, one of the 10 subsets is selected as the testing dataset while the remaining

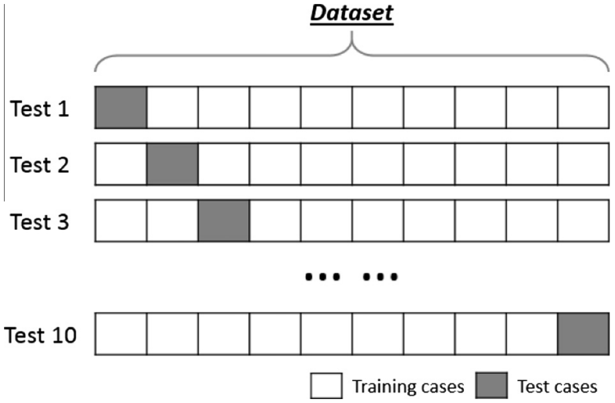

Fig. 2. 5-Fold cross validation.

9 subsets form the training dataset. The average accuracy of the 10 rounds of testing is calculated as the eventual prediction accuracy. In this case, each subset is selected as the testing dataset once, which makes the bias smaller.

By applying cross-validation, we are able to do a more objective comparative study on the feature selection strategies and the identification of regression algorithms. In addition, in order to reach the full potential of the regression algorithms, the optimal parameters of each algorithm should be identified. A common and efficient way to do so is named the grid search [40], which is simply an exhaustive searching through manually specified subsets of the parameter space of the algorithm. In practice, as shown in Fig. 3, usually two iterations of grid search helps to identify a pair of parameters that is close to the optimal value [20].

2.4. Regression

The methodology using multiple features or predictors to determine the energy consumption is named regression. One of the most typical regression algorithms is the multiple linear regression (MLR), the coefficients of which are usually identified using least squares [41]. In order to address the problem of overfitting, researchers have developed many shrinkage methods for MLR. Shrinkage methods are a general technique to improve a least-squares estimator which consists in reducing the variance by adding constraints on the value of coefficients [37]. Commonly seen methods in statistics [37] include the ridge regression (Ridge) [42], least absolute shrinkage and selection operator (LASSO) [43], and Elastic Net [44]. They have also been utilized by several energy related studies [33,45], and yielded good performance and fast computation.

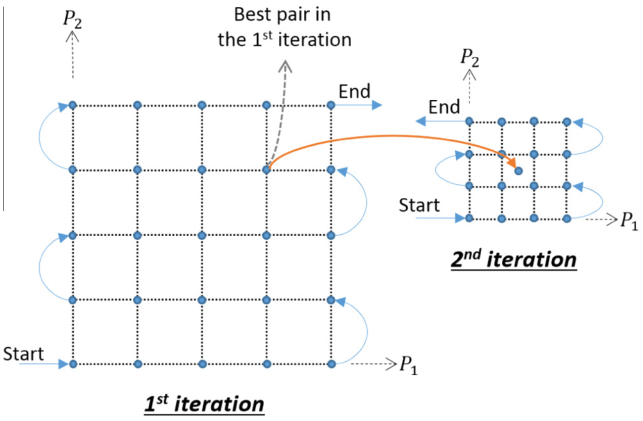

Fig. 3. Example of two iteration grid search. The 2nd iteration is grid searching the surrounding values of the best pair from the 1st iteration using smaller grids.

In addition, many machine learning algorithms nowadays are attracting increasing attention due to their powerful prediction capability in non-linear problems [2,22,25,32]. According to the literature, the two most commonly used machine learning algorithms in energy related fields are ANN [22,25,30] and SVR [2,31,32]. Considering different problems may have different choices of the optimal algorithm, this study will therefore compare the performance of these representative algorithms under different feature selection strategies.

2.4.1. Elastic Net

Elastic Net, which was firstly mentioned by Zou and Hastie [44], is a penalized shrinkage method that integrates LASSO and Ridge regression. It defines the Elastic Net criterion, the likelihood function, in Eq. (2).

$$L(\lambda_1, \lambda_2, \beta) = \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1 \tag{2}$$

where $\boldsymbol{y}$ is dependent response, $\boldsymbol{X}$ is the vector of features or variables, β is the vector of the regression coefficients in the linear regression, and λ_1, λ_2 are fixed non-negative parameters. The coefficients β is estimated by the minimizer of Eq. (2). The procedure can be viewed as a penalized least squares method, and the solution is equivalent to the optimization problem in Eq. (3) [44].

$$\begin{aligned} &\min_\beta \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 \\ s.t.\ &\alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1 \leqslant t \end{aligned} \tag{3}$$

where $\alpha = \lambda_2(\lambda_1 + \lambda_2)$, t is a constant in the optimization process. $\alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1$ is called as the Elastic Net penalty, which is a convex combination of the LASSO and ridge penalty. When $\alpha = 1$ it becomes the simple ridge regression, and when $\alpha = 0$ it becomes the LASSO penalty [19,44]. The R package, Glmnet, developed by Hastie and Qian [46], reformulate the objective function of Elastic Net for the Gaussian family distribution into Eq. (4).

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2N}\sum_{i=1}^{N}(y_i - x_i\beta)^2 + \lambda[(1-\alpha)\|\beta\|^2/2 + \alpha\|\beta\|_1] \tag{4}$$

where N is the number of training cases, $\lambda > 0$ is the complexity parameter. Coordinate descent can be applied to solve the problem [47]. Features with non-zero coefficients β after the optimization are viewed as selected features. According to Hastie and Qian [46], when training Elastic Net models, the complexity parameter λ and the penalty factor α need to be tuned, and usually a more stable optimal λ is given by the largest value which makes the cross-validation error within one standard error of the minimum error.

2.4.2. Artificial Neural Network

ANNs are computational models inspired by the behavior of neurons and the electrical signals they convey between input, processing, and output in a brain [48]. The differences in how neurons semantically communicate in a model results in different types of ANN. Examples include the feed-forward neural network, the radial basis function network, and neural networks in deep learning [48,49]. The feed-forward neural network is the most typical ANN, and with proper training, it is flexible enough to approximate any smooth function [38]. In addition, it has been implemented by many energy related studies and yielded good performance [22,25,30,50], so it is selected in the comparative study in this paper.

A generic feed-forward neural network is shown in Fig. 4. The output unit is calculated using Eq. (5).

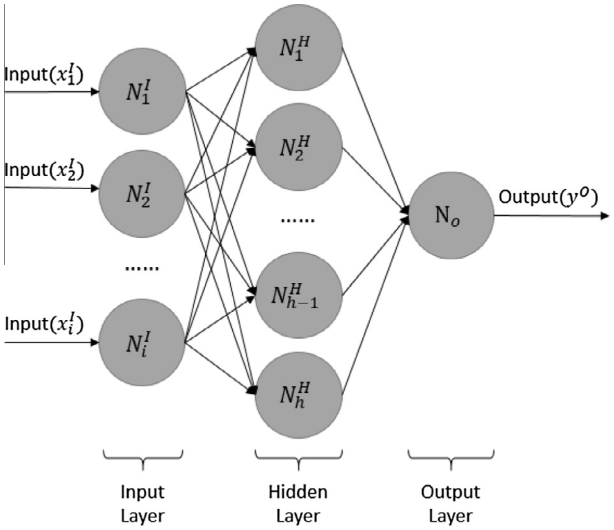

Fig. 4. A generic feed-forward neural network.

$$y^O = \varphi^O\left(\alpha^O + \sum_h \omega_h^H \varphi_h^H\left(\alpha_h^H + \sum_i \omega_i^I x_i^I\right)\right) \tag{5}$$

where y is the output of the model, φ is the activation function, α is the constant bias, ω is the weight of each unit, and O, H, I refers to the output layer, hidden layer and input layer, respectively. The activation function φ_h^H in the hidden layer is almost always taken to be the logistic function, while the output function φ^O is a linear function in regression [38].

Identification of the weights and biases is solved by minimizing the loss function (e.g., least squares) with the regularization shown in Eq. (6) using gradient descent [38].

$$L(y) = \sum_k |y_k^O - y_k|^2 + \lambda_d \boldsymbol{\omega}^2 \tag{6}$$

where k refers to the training cases, y_k is the true value of case k, $\boldsymbol{\omega}^2$ is the sum of squares of the weights, and λ_d is the weight decay, the use of which seems both to help the optimization process and to avoid over-fitting [38,51]. According to Venables and Ripley [38], the size of the hidden layer h and the weight decay λ_d are sensitive to the model performance and need to be tuned in practice.

2.4.3. Support Vector Regression

SVR is the version of Support Vector Machine (SVM) for regression. The goal of SVR [52] is to find a function f(x) that has at most ε deviation from the actually obtained targets y_i for all the training data, and at the same time is as flat as possible. In other words, we do not care about errors as long as they are less than ε, but would not like to accept any deviation larger than this. Such a concept of learning yielded good performance in many energy related studies [2,31,32]. The algorithm tries to solve the convex optimization problem shown in Eq. (7).

$$\begin{array}{ll} \min & \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\ s.t. & \begin{cases} y_i - \boldsymbol{\omega x} - b \leqslant \varepsilon + \xi_i \\ y_i - \boldsymbol{\omega x} - b \geqslant -\varepsilon - \xi_i^* \\ \xi_i, \xi_i^* \geqslant 0 \end{cases} \end{array} \tag{7}$$

where $\boldsymbol{\omega}$ is the vector of the regression coefficients, b is the regression intercept, ε is the deviation margin, C is a regularization term that determines the degree of the linear penalty applied to the residual excess. ξ_i and ξ_i^*, which are illustrated in Fig. 5, are the slack variables in the "soft margin" loss function [52].

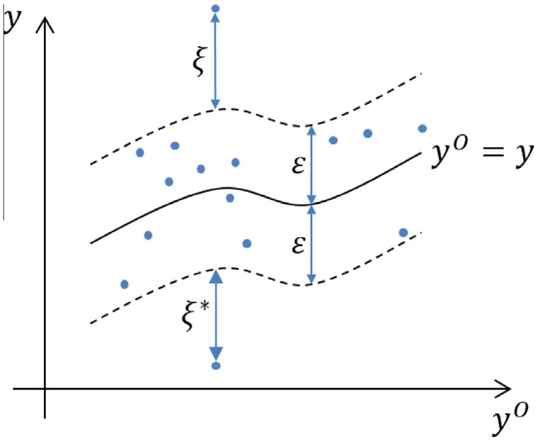

Fig. 5. Illustration of ξ and ξ^* in the hyper-plane.

The optimization problem in Eq. (7) can be solved using a standard dualization method utilizing Lagrange multipliers [52], and the solution is given by Eq. (8).

$$y^O = \sum_{i=1}^{n_{sv}}(\alpha_i - \alpha_i^*)K(\boldsymbol{x}', \boldsymbol{x}) \tag{8}$$

where n_{sv} is the number of support vectors, α_i and α_i^* are Lagrange multipliers, and $K(\boldsymbol{x}', \boldsymbol{x})$ is the kernel function. The Gaussian radial basis function (RBF) shown in Eq. (9) is used as the kernel in this study due to its ability in generalizing non-linear functions and its efficiency in large datasets [2].

$$K(\boldsymbol{x}', \boldsymbol{x}) = \exp(-\gamma|\boldsymbol{x} - \boldsymbol{x}'|^2) \tag{9}$$

where γ is the kernel parameter that defines the radius of influence for each data point. According to Jain et al. [2], the kernel parameter γ, the training tolerance ε and the regularization term C are sensitive to the model performance and need to be tuned in practice.

3. Estimation of the energy use intensity of residential buildings in New York City

To test the effectiveness of the proposed methodology framework and compare the different feature selection strategies and regression algorithms, we conducted a case study in New York City (NYC) by estimating the site energy use intensity (EUI) of multifamily residential buildings. The reasons to choose NYC include, firstly, there is existing literature [6,19] estimating the building energy consumption in NYC, which allows a comparison with our results. Second, the government of NYC has publicized more than 1000 datasets [53] regarding various aspects of city performance such as education, housing and transportation. This allows the application of our methodology framework.

3.1. Data collection and preprocessing based on GIS

Due to the Local Law 84 of the NYC government [54], the general public is able to get both of the site EUI and source EUI data of buildings in NYC. In order to make the results comparable, the site EUI was selected as the target in the regression model following Hsu's work [19]. The latest available data is for the year of 2013 [53]. The parcel number system, Borough Block and Lot (BBL), is provided to identify the buildings. By connect the BBL to the building data in the primary land use tax lot output (PLUTO) database provided by the Department of City Planning (DCP), we were able to get the data on detailed areas for different building usage. To avoid the influence from mixed residential/commercial buildings, only buildings with 100% residential areas were selected in the study. This resulted in 3854 multi-family residential buildings. After removing the outliers using boxplot statistics, there remained 3640 buildings, comprised of 1097 buildings in Brooklyn, 933 in Manhattan, 853 in The Bronx, 725 in Queens, and 32 in Staten Island.

The raw data of the features were mainly collected from three data sources: (1) the PLUTO database provided by DCP, (2) the 2013 American Community Survey (ACS) 5-year estimates from the U.S. Census Bureau, and (3) the NYC Open Data [53]. After the GIS integrated preprocessing, there were 216 continuous features prepared for the regression. Table 1 summarizes the source of these features with different categories.

The GIS tool used in this study was ArcGIS 10.2, and it played an important role during the preprocessing. First is the data integration, which was done with the help of the geocoding toolbox in ArcGIS, and the shapefiles from NYC department of city planning and the government website of NY state. Second is the missing data. By using the geostatistical analyst tool in ArcGIS, some missing values in the ACS database were filled with the multiple linear regression value based on the geolocations. An example is shown in Fig. 6. Last is feature generation. As shown in Table 1, there are 64 features in total generated in ArcGIS. The majority of the 64 features are generated using the shapefiles or shape areas calculated in ArcGIS to get the density value, such as population density and traffic density; or using the spatial analyst tool to calculate the distance to the nearest place of interest, such as the distance to the coast line, and distance to the nearest subway entrance. One special example is the generation of the vegetation feature. The reason why this feature is worthy of studying is explained in Section 2.1. The feature was added based on the Normalized Difference Vegetation Index (NDVI), which is an index of plant "greenness" or photosynthetic activity, and is one of the most commonly used vegetation indices [45]. As shown in Fig. 7, by downloading the remote satellite sensing figures captured by the Landsat 8 satellite from the U.S. Geological Survey, ArcGIS is able to calculate the NDVI [55,56]. The scene shown in Fig. 7 was captured at 15:33 pm July 31th, 2014, with 2% cloud coverage.

3.2. Results and discussion

The programs are coded in R x64 3.2.2 using a PC with an i7-3770 CPU at 3.40 GHz and Windows 7 Enterprise 64-bit OS installed. In order to make the results comparable, the features and the target were standardized and transformed following Hsu's work [19]. The calculation results, which are measured by the mean squared error (MSE), are shown in Table 2. Due to the GIS integrated feature engineering process, the optimal MSE (0.838) of the Elastic Net in this study is lower than the MSE in Hsu's work [19].

The last column of Table 2 shows the average MSE of the algorithms using different feature selection strategies. It can be seen that Elastic Net has the highest MSE with 0.838 compared to ANN and SVR. This is because although Elastic Net uses shrinkage methods to regularize the estimators, the algorithm is still based on linear models, while ANN and SVR were designed to model non-linear regressions. On the other hand, compared to the other two algorithms, Elastic Net only took 0.9 s to model the regression, and the MSE is only 0.013 higher than ANN using the full set of features. Such an efficient performance is the reason why this algorithm is widely used in computer engineering [38]. SVR, with an average MSE of 0.750, outperforms the ANN (0.802), and the computation time is significantly less than ANN, regardless of the feature selection methods. This shows that, to the regression problem in this study, SVR is a better choice.

The characteristics of different feature selection methods can also be reflected in Table 2. First is the filter method. By comparing the results calculated using the full set of features (Null) and the CFS filter method, it can be seen that CFS helped to significantly reduce the feature size and the computation time, while did not lose much MSE. Second is the wrapper method. Greedy forward selection helped produce a lower MSE than either CFS or Null, and the number of selected features is even smaller than in the CFS method. However, the drawback of this method is quite obvious. Its computation time is dramatically longer than the other methods. Last is the embedded method. The performance of Elastic Net is limited by the linear model as mentioned previously, but the features selected by Elastic Net may be useful. In fact, this study tried building SVR and ANN models on the 45 features selected by Elastic Net, and to our surprise, such a combination produced good performances. For example, the model generated by SVR + Elastic Net not only had the lowest MSE (0.728) but also a very short computation time, which is only 1.7 s slower than SVR + CFS.

As a result, the best-performance model generated by SVR + Elastic Net was used to estimate the site EUI of the rest of the multi-family residential buildings in NYC. According to the PLUTO database [57], there are totally 143,080 multi-family residential buildings in NYC. By excluding the 3640 buildings used previously, there were 139,440 left. The 216 features of these 139,440 buildings were extracted and prepared from the same data source using the same GIS integrated process. Fig. 8 shows a map of the median estimated site EUI of the multi-family residential buildings on a Block Group basis, which is the minimum survey unit in American census. It can be seen that the Bronx, southeastern Queens, and northern Manhattan have relatively higher site EUI, while Staten Island, southern Manhattan and southern Brooklyn have relatively lower site EUI. Such a map can facilitate city planning during policy making. For example, districts with higher median site EUI reveal higher necessity for building retrofitting, and the government is suggested to put out relative policies to improve the energy performance of those districts.

Table 3 shows the 45 features selected and used in the SVR + Elastic Net model. It can be seen that 19 out of the selected 45

Table 1
Summary of the prepared features in different categories.

Category	ACS	ArcGIS	PLUTO	Open data	Total
Building	10	4	15	2	31
Demography	14	7		1	22
Economy	24	14	10		48
Education	12	5			17
Environment		4		5	9
Households	48	18			66
Surrounding		12		3	15
Transportation	4			4	8
Total	112	64	25	15	216

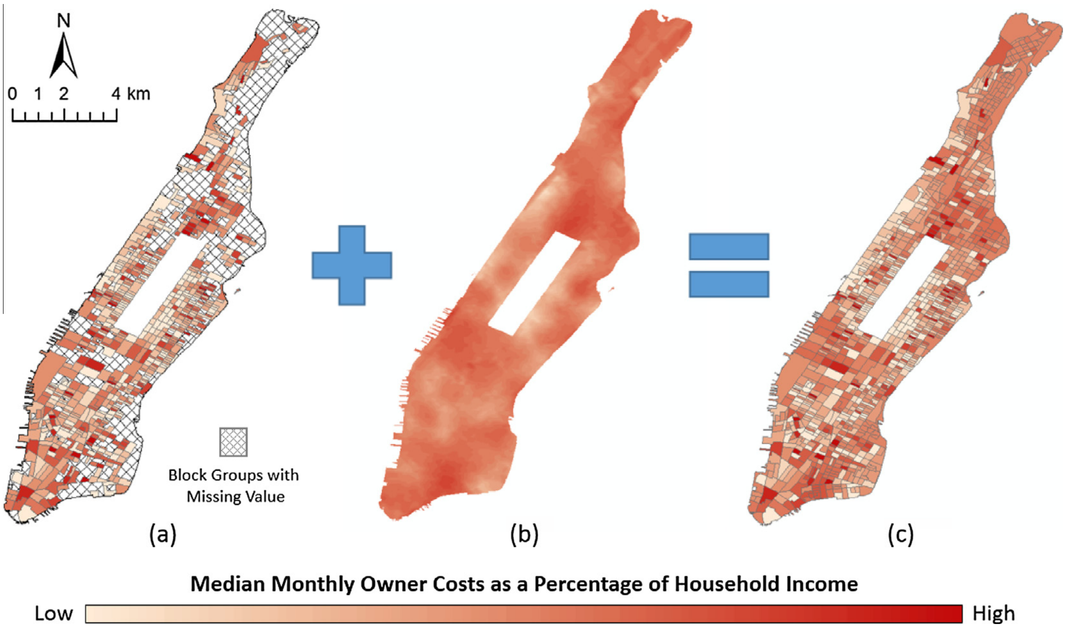
N
0 1 2 4 km
Block Groups with Missing Value
(a)
(b)
(c)
Median Monthly Owner Costs as a Percentage of Household Income
Low
High

Fig. 6. An example of filling missing values using ArcGIS. For the presentation purpose, the map was cropped into Manhattan borough. (a) The plot of median monthly owner cost as a percentage of household income from the ACS datasource based on Block Groups with missing values. (b) The estimated value distribution using the geostatistical analyst toolbox in ArcGIS. (c) The plot with missing values filled.

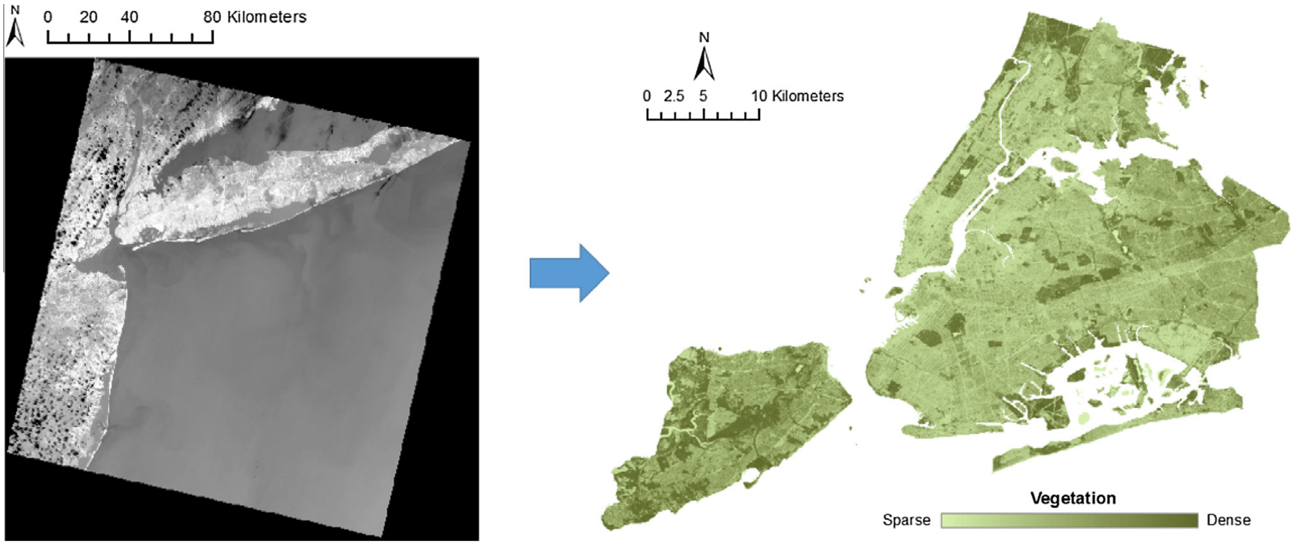

Fig. 7. Calculation of NDVI from the Landsat 8 satellite figures using ArcGIS.

Table 2
Summary of the MSEs calculated by using different feature selection strategies and regression algorithms.

Algorithm	Feature selection strategy	Number of selected features	Cross-validation MSE (10-fold)	Computation time (s)	Average MSE
SVR	Null	216	0.759	43.4	0.750
	CFS	25	0.773	14.0	
	Greedy forward	19	0.739	11.2 h	
	Elastic Net	45	0.728	15.7	
ANN	Null	216	0.825	815.6	0.802
	CFS	25	0.832	39.4	
	Greedy forward	17	0.781	13.6 h	
	Elastic Net	45	0.770	61.8	
Elastic Net	Elastic Net	45	0.838	0.9	0.838
	LASSO ($\alpha = 1$)[a]	N/A	0.990	N/A	0.993
	Ridge ($\alpha = 0$)[a]	N/A	1.000	N/A	
	Elastic Net[a]	N/A	0.990	N/A	

[a] The results from Hsu's work [19], and N/A means not available.

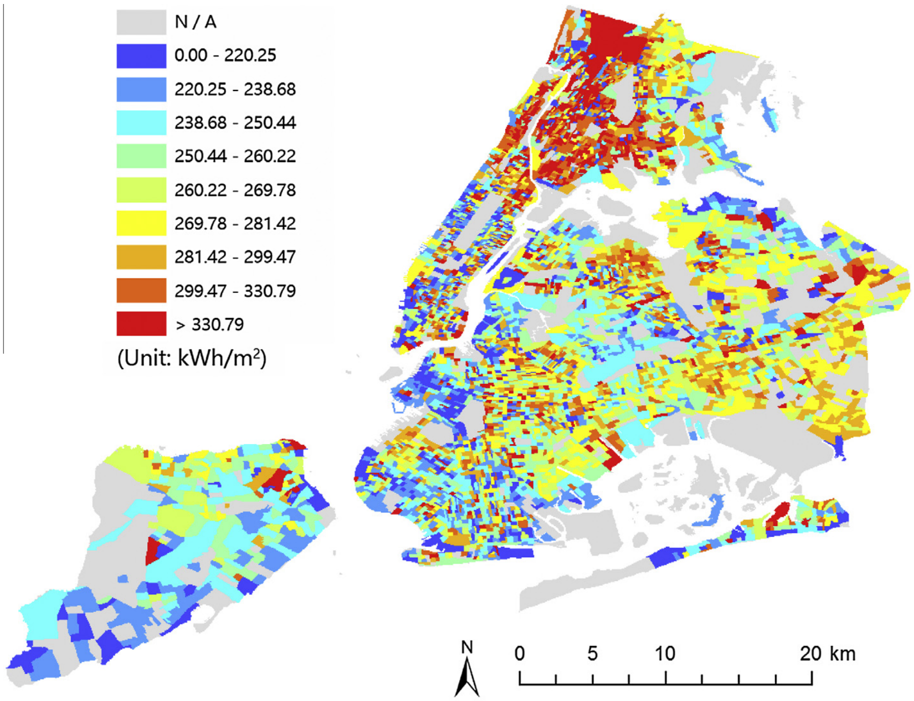

Fig. 8. The map of the median estimated site EUI of the multi-family residential buildings in NYC on a Block Group basis.

Table 3
Summary of the selected 45 features by Elastic Net.

Category	Features
Building (11)	Building age; total residential units; num. of floors; mean residential unit size; lot frontage; nearby complaints about residential buildings[b]; floor area ratio; building frontage; % households heat house by gas[a]; % households heat house by fuel[a]; median building age of owner occupied households[a,b]
Demography (3)	Population density[a,b]; density of population over 60[a,b]; % white people[a]
Economy (10)	Median gross rent as % of household income[a,b]; median gross rent of occupied units paying rent[a,b]; median households income[a,b]; median monthly owner costs[a,b]; median monthly owner costs as % of household income[a,b]; median monthly owner costs of housing units with a mortgage[a,b]; unemployment[a]; exempt land value; % exempt land value over total land value; % exempt total value over total property value
Education (3)	% 25 or over that has bachelor degree or higher[a]; % people without any education[a]; num. of people currently enrolled in college or graduated school[a]
Environment (2)	Avg. nearby NDVI[b]; nearby complaints about party noise[b]
Households (8)	Avg. household size[a,b]; % 1-person households[a]; % family households[a]; % households living in buildings built after 2000[a]; % households living in buildings built before 1980[a]; % households that has people over 60[a]; % people go to work in the afternoon[a]; % vacant housing units[a]
Surrounding (5)	Num. of nearby senior centers[b]; num. of nearby subway entrances[b]; dist. to the nearest pedestrian plaza[b]; num. of nearby hospitals or clinics[b]; dist. to the nearest subway entrance[b]
Transportation (3)	Mean travel time to work[a,b]; % people drive alone to work[a]; % people using public transportation to work[a]

[a] Features that were prepared from the ACS database, are based on the Block Group where the building locates.

[b] Features that were either using GIS to fill the missing values or generated with the help of GIS.

are related to GIS, and this, from another aspect, supports the effectiveness of integrating GIS in feature engineering. In addition, as shown in Table 3, the categories of building, economy and households have more selected features, while the categories of demography, education, environment and transportation have less selected features. However, this study did not further rank and evaluate the feature influence of each of these 45 features, because of two reasons. Firstly, the exploration of feature influence should be based on the SVR model, and since in SVR models the regression is based on support vectors, the feature influence or coefficients of the features cannot be directly examined. The related literatures in machine learning are more about feature selection using the wrapper methods instead of evaluating the variable importance [58,59]. Secondly, 171 features are not selected by Elastic Net, but it does not mean that they are all less important or irrelevant to the site EUI of the buildings. For example, one feature is the year when the latest major renovation of the building was conducted, and if no renovation ever happened it will use the year when the building was built. This feature has a high correlation with the building age, and is expected to be related to the site EUI of the buildings. However, it was not selected by Elastic Net. On the other hand, it is hard to explain the relationships between the site EUI and a few of the selected features, such as the number of nearby hospitals in the surrounding category. They were shortlisted either because they might be the least influential features in the selected 45, or were included due to the data noise. A further systematic methodology should be designed to evaluate the feature influence of all these 216 features over the site EUI in a more objective way.

To further support the effectiveness of the proposed methodology framework, the source EUI of the 3640 residential buildings is also estimated to compare with the results in the study conducted by Howard et al. [6]. In their work, floor areas and regional energy consumption were used to estimate the source EUI of different building types. The comparison of the average estimated source EUI of residential buildings in each borough is shown in Table 4. The results of this study were calculated through a 10-fold cross-estimation process using SVR + Elastic Net. It can be seen that, if

Table 4
Comparison of the mean estimated source EUI in each borough in NYC with the past work.

Borough	Average reported source EUI from the government data in 2013 (kW h/m^2)	Source EUI estimated by Howard et al. [6]		Source EUI estimated in this study	
		Average	Error	Average	Error
Brooklyn	367.05	356.70	2.82%	371.92	1.33%
The Bronx	400.35	277.70	30.64%	406.00	1.41%
Manhattan	383.27	311.90	18.62%	385.74	0.64%
Queens	375.74	356.70	5.07%	382.99	1.93%
Staten Island	412.83	356.70	13.60%	429.85	4.12%
Average	387.85	331.94	14.15%	395.30	1.89%

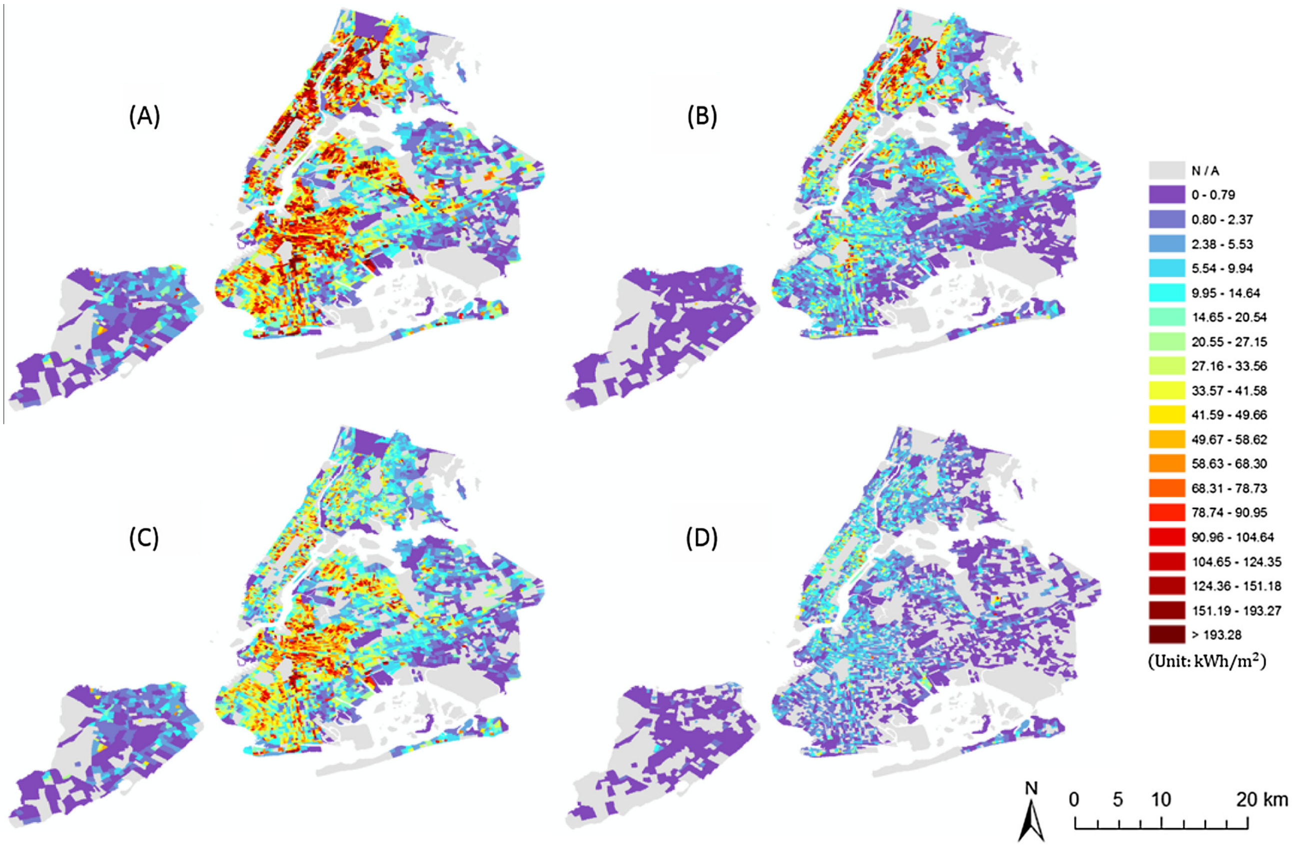

Fig. 9. Estimated energy demand per Block Group area for space heating of multi-family residential buildings in NYC. (A) Total energy demand for space heating. (B) Fuel demand for space heating. (C) Gas demand for space heating. (D) Electric demand for space heating.

using the average self-reported source EUI from the government data in 2013 [54] as the benchmark, the average error of our estimations is 1.89%, which is lower than that estimated by Howard et al. [6]. In addition, by utilizing the data on the annual energy consumed for different end use of residential buildings in NY state from the Residential Energy Consumption Survey [6,60], the total energy demand for different end use of the multi-family residential buildings can also be estimated. An example of the estimated total energy demand for space heating is shown in Fig. 9. Since the energy demand for space heating is quite seasonal, the distribution shown in Fig. 9 can be useful reference when planning urban energy infrastructures.

4. Conclusions

To conclude, this study proposes a methodology framework to estimate the building energy use intensity on the urban scale by integrating GIS and Big Data technology. The framework addressed the major challenges of data mining in the urban scale, including preprocessing, feature selection, and algorithm optimization. A case study on estimating the energy use intensity of 3640 multi-family residential buildings in NYC was conducted to test the effectiveness of the proposed methodology framework. The results showed that the framework was able to help produce lower estimation error than in previous research.

The contributions of the work can be summarized into three aspects. First is the methodology framework. Feature engineering is known as the key in data mining [20]. The integration of GIS into this most important data mining step has shown lots of benefits, such as connecting high dimension geo based datasets conveniently, filling the missing values nicely, and helping generate many useful features which may be easily overlooked using traditional feature management tools in urban scale energy related problems. In addition, data mining is such a practical tool that by changing the features, the target, and the algorithms, the framework can be easily extended to solve other urban scale research problems, such as estimating the urban scale air quality distribution and the noise pollution. Second is the case study on the residential buildings in NYC. By using the extensive open data in NYC, the study was able to estimate the energy use intensity of residential buildings with less error than previous research, and the estimated maps can be useful references for energy planning and policy making. Last is the comparative study on feature selection strategies and regression models. The case study shows that the filter methods can significantly reduce the feature size and the computation time, and some even produce better performance (filter features using Elastic Net). The wrapper methods generally have better performance than using the full set of features but dramatically increase the computation time. The comparison between different algorithms shows that SVR is an efficient and effective regression algorithm for non-linear models. ANN, particularly the feed-forward neural network, underperformed SVR and consumed longer computation time according to the case study. However, ANN is still believed to be a worth-trying regression tool for non-linear problems considering its extensive possibilities for further adjustment.

The limitation of this paper is that the data used in the case study are not balanced. Only 32 out of the 3640 buildings are from Staten Island, which makes the average prediction error for that borough higher than the others in Table 4. Future work can be to explore a systematic methodology to objectively evaluate how the features influence the energy use intensity.

Acknowledgement

The authors would like to acknowledge the support by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 611513).

References

[1] Yuan J, Farnham C, Emura K. Development and application of a simple BEMS to measure energy consumption of buildings. Energy Build 2015;109:1–11. http://dx.doi.org/10.1016/j.enbuild.2015.10.012.

[2] Jain RK, Smith KM, Culligan PJ, Taylor JE. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. Appl Energy 2014;123:168–78. http://dx.doi.org/10.1016/j.apenergy.2014.02.057.

[3] Petersen S, Svendsen S. Method and simulation program informed decisions in the early stages of building design. Energy Build 2010;42:1113–9. http://dx.doi.org/10.1016/j.enbuild.2010.02.002.

[4] Virote J, Neves-Silva R. Stochastic models for building energy prediction based on occupant behavior assessment. Energy Build 2012;53:183–93. http://dx.doi.org/10.1016/j.enbuild.2012.06.001.

[5] Kusiak A, Xu G. Modeling and optimization of HVAC systems using a dynamic neural network. Energy 2012;42:241–50. http://dx.doi.org/10.1016/j.energy.2012.03.063.

[6] Howard B, Parshall L, Thompson J, Hammer S, Dickinson J, Modi V. Spatial distribution of urban building energy consumption by end use. Energy Build 2012;45:141–51. http://dx.doi.org/10.1016/j.enbuild.2011.10.061.

[7] Huang K-T, Hwang R-L. Future trends of residential building cooling energy and passive adaptation measures to counteract climate change: The case of Taiwan. Appl Energy 2016. http://dx.doi.org/10.1016/j.apenergy.2015.11.008.

[8] Gruber M, Trüschel A, Dalenbäck J-O. Energy efficient climate control in office buildings without giving up implementability. Appl Energy 2015;154:934–43. http://dx.doi.org/10.1016/j.apenergy.2015.05.075.

[9] Zhou Y, Clarke L, Eom J, Kyle P, Patel P, Kim SH, et al. Modeling the effect of climate change on U.S. state-level buildings energy demands in an integrated assessment framework. Appl Energy 2014;113:1077–88. http://dx.doi.org/10.1016/j.apenergy.2013.08.034.

[10] Ye H, He X, Song Y, Li X, Zhang G, Lin T, et al. A sustainable urban form: the challenges of compactness from the viewpoint of energy consumption and carbon emission. Energy Build 2015;93:90–8. http://dx.doi.org/10.1016/j.enbuild.2015.02.011.

[11] Martins TAL, Adolphe L, Bastos LEG. From solar constraints to urban design opportunities: optimization of built form typologies in a Brazilian tropical city. Energy Build 2014;76:43–56. http://dx.doi.org/10.1016/j.enbuild.2014.02.056.

[12] Vartholomaios A. The residential solar block envelope: a method for enabling the development of compact urban blocks with high passive solar potential. Energy Build 2015;99:303–12. http://dx.doi.org/10.1016/j.enbuild.2015.04.046.

[13] Liu J, Heidarinejad M, Gracik S, Srebric J. The impact of exterior surface convective heat transfer coefficients on the building energy consumption in urban neighborhoods with different plan area densities. Energy Build 2015;86:449–63. http://dx.doi.org/10.1016/j.enbuild.2014.10.062.

[14] Gaigné C, Riou S, Thisse J-F. Are compact cities environmentally friendly? J Urban Econ 2012;72:123–36. http://dx.doi.org/10.1016/j.jue.2012.04.001.

[15] Shammin MR, Herendeen RA, Hanson MJ, Wilson EJH. A multivariate analysis of the energy intensity of sprawl versus compact living in the U.S. for 2003. Ecol Econ 2010;69:2363–73. http://dx.doi.org/10.1016/j.ecolecon.2010.07.003.

[16] Norman J, MacLean H, Kennedy C. Comparing high and low residential density: life-cycle analysis of energy use and greenhouse gas emissions. J Urban Plan Dev 2006;132:10–21. http://dx.doi.org/10.1061/(ASCE)0733-9488(2006)132:1(10).

[17] Bradshaw JL, Bou-Zeid E, Harris RH. Comparing the effectiveness of weatherization treatments for low-income, American, urban housing stocks in different climates. Energy Build 2014;69:535–43. http://dx.doi.org/10.1016/j.enbuild.2013.11.035.

[18] Santamouris M, Alevizos SM, Aslanoglou L, Mantzios D, Milonas P, Sarelli I, et al. Freezing the poor—Indoor environmental quality in low and very low income households during the winter period in Athens. Energy Build 2014;70:61–70. http://dx.doi.org/10.1016/j.enbuild.2013.11.074.

[19] Hsu D. Identifying key variables and interactions in statistical models of building energy consumption using regularization. Energy 2015;83:144–55. http://dx.doi.org/10.1016/j.energy.2015.02.008.

[20] Han J, Kamber M, Pei J. Data mining: concepts and techniques. Elsevier; 2011.

[21] Koo C, Hong T. Development of a dynamic operational rating system in energy performance certificates for existing buildings: geostatistical approach and data-mining technique. Appl Energy 2015;154:254–70. http://dx.doi.org/10.1016/j.apenergy.2015.05.003.

[22] Siami-Irdemoosa E, Dindarloo SR. Prediction of fuel consumption of mining dump trucks: a neural networks approach. Appl Energy 2015;151:77–84. http://dx.doi.org/10.1016/j.apenergy.2015.04.064.

[23] Long H, Zhang Z, Su Y. Analysis of daily solar power prediction with data-driven approaches. Appl Energy 2014;126:29–37. http://dx.doi.org/10.1016/j.apenergy.2014.03.084.

[24] Cheng JCP, Ma LJ. A data-driven study of important climate factors on the achievement of LEED-EB credits. Build Environ 2015;90:232–44. http://dx.doi.org/10.1016/j.buildenv.2014.11.029.

[25] Cheng JCP, Ma LJ. A non-linear case-based reasoning approach for retrieval of similar cases and selection of target credits in LEED projects. Build Environ 2015;93(Part 2):349–61. http://dx.doi.org/10.1016/j.buildenv.2015.07.01.

[26] Geographic information system. Wikipedia Free Encycl; 2015.

[27] De Gennaro M, Paffumi E, Scholz H, Martini G. GIS-driven analysis of e-mobility in urban areas: an evaluation of the impact on the electric energy grid. Appl Energy 2014;124:94–116. http://dx.doi.org/10.1016/j.apenergy.2014.03.003.

[28] Sánchez-García S, Canga E, Tolosana E, Majada J. A spatial analysis of woodfuel based on WISDOM GIS methodology: multiscale approach in Northern Spain. Appl Energy 2015;144:193–203. http://dx.doi.org/10.1016/j.apenergy.2015.01.099.

[29] Xu J, Song X, Wu Y, Zeng Z. GIS-modelling based coal-fired power plant site identification and selection. Appl Energy 2015;159:520–39. http://dx.doi.org/10.1016/j.apenergy.2015.09.008.

[30] Carta JA, Cabrera P, Matías JM, Castellano F. Comparison of feature selection methods using ANNs in MCP-wind speed methods. A case study. Appl Energy 2015;158:490–507. http://dx.doi.org/10.1016/j.apenergy.2015.08.102.

[31] Majidpour M, Qiu C, Chu P, Pota HR, Gadh R. Forecasting the EV charging load based on customer profile or station measurement? Appl Energy 2016;163:134–41. http://dx.doi.org/10.1016/j.apenergy.2015.10.184.

[32] Patil MA, Tagade P, Hariharan KS, Kolake SM, Song T, Yeo T, et al. A novel multistage Support Vector Machine based approach for Li ion battery remaining useful life estimation. Appl Energy 2015;159:285–97. http://dx.doi.org/10.1016/j.apenergy.2015.08.119.

[33] Huebner GM, Hamilton I, Chalabi Z, Shipworth D, Oreszczyn T. Explaining domestic energy consumption – the comparative contribution of building factors, socio-demographics, behaviours and attitudes. Appl Energy 2015;159:589–600. http://dx.doi.org/10.1016/j.apenergy.2015.09.028.

[34] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[35] Hall MA. Correlation-based feature selection for machine learning. The University of Waikato; 1999.

[36] Lal TN, Chapelle O, Weston J, Elisseeff A. Embedded methods. Feature Extr.: Springer; 2006. p. 137–65.

[37] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer; 2009.

[38] Venables WN, Ripley BD. Modern applied statistics with S-PLUS. Springer Science & Business Media; 2013.

[39] Amiri SS, Mottahedi M, Asadi S. Using multiple regression analysis to develop energy consumption indicators for commercial buildings in the U.S. Energy Build 2015;109:209–16. http://dx.doi.org/10.1016/j.enbuild.2015.09.073.

[40] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13:281–305.

[41] Regression analysis. Wikipedia Free Encycl; 2015.

[42] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12:55–67.

[43] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 1996:267–88.

[44] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 2005;67:301–20.

[45] Douak F, Melgani F, Benoudjit N. Kernel ridge regression with active learning for wind speed prediction. Appl Energy 2013;103:328–40. http://dx.doi.org/10.1016/j.apenergy.2012.09.055.

[46] Hastie T, Qian J. *Glmnet Vignette*. Retrieved from https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html.

[47] Friedman J, Hastie T, Tibshirani R. glmnet: Lasso and elastic-net regularized generalized linear models. R Package Version 2009:1.

[48] Haykin SS, Haykin SS, Haykin SS, Haykin SS. Neural networks and learning machines, vol. 3. Pearson Education Upper Saddle River; 2009.

[49] Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85–117. http://dx.doi.org/10.1016/j.neunet.2014.09.003.

[50] Yuce B, Rezgui Y, Mourshed M. ANN–GA smart appliance scheduling for optimized energy management in the domestic sector. Energy Build 2016;111:311–25. http://dx.doi.org/10.1016/j.enbuild.2015.11.017.

[51] Ripley BD. Pattern recognition and neural networks. Cambridge University Press; 1996.

[52] Smola A, Vapnik V. Support vector regression machines. Adv Neural Inf Process Syst 1997;9:155–61.

[53] NYC Open Data. NYC Open Data, n.d. https://data.cityofnewyork.us/ [accessed December 8, 2015].

[54] City of New York MO of S. Benchmarking data disclosure & reports; 2015.

[55] Sarkar C, Webster C, Gallacher J. Healthy cities: public health through urban planning. Edward Elgar Publishing; 2014.

[56] Roy DP, Wulder MA, Loveland TR, Woodcock CE, Allen RG, Anderson MC, et al. Landsat-8: Science and product vision for terrestrial global change research. Remote Sens Environ 2014;145:154–72. http://dx.doi.org/10.1016/j.rse.2014.02.001.

[57] City of New York D of CP. PLUTO and MapPLUTO; 2015.

[58] Rakotomamonjy A. Variable selection using svm based criteria. J Mach Learn Res 2003;3:1357–70.

[59] Zhang HH, Liu Y, Wu Y, Zhu J, et al. Variable selection for the multicategory SVM via adaptive sup-norm regularization. Electron J Stat 2008;2:149–67.

[60] No OMB. Residential energy consumption survey; 2005.