Energy & Buildings 204 (2019) 109519

Contents lists available at ScienceDirect

Energy & Buildings

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

Comparison of regression models for estimation of carbon emissions during building's lifecycle using designing factors: a case study of residential buildings in Tianjin, China

Mao Xikai[a], Wang Lixiong[a], Li Jiwei[a,b,*], Quan Xiaoli[a], Wu Tongyao[a]

[a] *School of Architecture, Tianjin University, Tianjin 300072, China*
[b] *College of Civil Engineering and Architecture, Hebei University, Baoding 071002, China*

ARTICLE INFO

Article history:
Received 20 May 2019
Revised 7 October 2019
Accepted 10 October 2019
Available online 11 October 2019

Keywords:
Building lifecycle
Carbon emissions
Predictive model
Residential building

ABSTRACT

Many studies have been conducted on life cycle assessment and control measures for carbon emissions of buildings. Methods proposed by these studies usually require not only specific accounting model, but also detailed inventory data, which is not available at early design stage. Seeing that the importance of design phase to carbon emissions during building's lifecycle, a study on regression model of carbon emissions using designing factors was done. Firstly, based on process analysis method, the carbon emissions of 207 residential buildings in Tianjin were calculated. The results show that annual carbon emissions per floor area are between 30 and 60 $kgCO_2/(m^2\bullet year)$, with manufacture phase and operation phase accounting for 11%–25% and 75%–87%, respectively. Then, correlation analysis and elastic net were used to determine 12 designing factors for predictive model; At last, four regression techniques, PCR, RF, MLP and SVR were used to develop regression models, respectively; comparison and process analysis of model development were given later. The results show that SVR has the optimal predictive accuracy among four models, its corresponding coefficient of determination can reach to 0.800. This regression model can be utilized to estimate carbon emissions based on designing factors, which can help designers make a strategic decision at early stage.

1. Introduction

Human activities are estimated to have caused approximately 1.0 °C of global warming above pre-industrial levels, and global warming is likely to reach 1.5 °C between 2030 and 2052 if it continues to increase at the current rate. Climate-related risks to health, livelihoods, food security, water supply, human security, and economic growth are projected to increase with global warming of 1.5 °C [1]. Against such background, it has become a major concern of the world, particularly of China to save energy and reduce greenhouse gas (GHG) emissions represented by CO_2 [2]. Building sector accounted for nearly 40% of the world's energy consumption, and 33% of the related GHG emissions [3]. In 2014, China's construction industry related carbon emissions reached 3.8 billion tons, exceeded one-third of the total [4]. In response to this, many attempts have initiated to find effective measures to lower building carbon emissions throughout the life cycle [5].

Life Cycle Assessment (LCA) is a process whereby the material and energy flow a system are quantified and evaluated. LCA allows for an evaluation of impacts of different processes and life cycle stages on the environment [6]. Many studies have used LCA methods to evaluate the life-cycle carbon emissions of buildings. Suzuki and Oka (1998) used Input/Output (I/O) to calculate the total domestic product, and estimated energy consumption and CO_2 emission during the entire life cycle of buildings [7]. Su et al. (2008) developed a life-cycle inventory model for the office buildings, and compared environmental effects of two different building structures (steel and concrete) [8]. Wu et al. (2012) used a process-based LCA to identify and quantify the energy consumption and CO_2 emissions of an office building in China [9]. Stephen and Crawford (2013) analyzed the total life cycle energy demand of a typical Belgian passive house, comprising embodied, operational and transport energy [10]. Zhang and Wang (2015) proposed a detailed carbon emission inventory, an analytical framework and evaluation indices, the proposed methodology was applied to three case buildings [11]. Besides the lifecycle, many studies also towards the estimation of carbon emissions during embodied or construction phase [5,12,13].

[*] Corresponding author at: School of Architecture, Tianjin University, Tianjin 300072, China.
E-mail address: 47887840@qq.com (L. Jiwei).

https://doi.org/10.1016/j.enbuild.2019.109519
0378-7788/© 2019 Elsevier B.V. All rights reserved.

Table 1
Approaches to data acquisition during different phases in studies.

Study	Location	Building type	Phases Manufacture	Erection	Operation	Disposal
Peng [22]	Nanjing, China	Office	BIM		Ecotect	Estimated value
Li et al. [28]	Nanjing, China	Residential	Glodon (software for quantity survey)		Reference value	Estimated value
Ma et al. [21]	Tianjin, China	Office	BOQ		Operational data	Excluded
Zhang et al. [33]	Tianjin, China	Residential	Construction entity, etc.		Energyplus	Estimated value
Ou et al. [27]	Huaian, China	Residential	BIM		Green Building Studio	Not clear
Wu et al. [31]	China	Residential & Commercial	Final report	Predicted	DeST	Predicted
Gustavsson et al. [30]	Sweden	Apartment	BOQ	Reference value	ENORM, ENSYST	Reference value
Basbagill et al. [25]	Confidential	Residential	BIM	Excluded	eQUEST	Excluded
Rosselló-Batle [26]	Balearic Islands, Spain	Hotel	TCQ Tools		Operational data	Approximation
Wallhagen et al. [29]	Gävle, Sweden	Office	Drawings & Specifications	Excluded	Enorm 1000	Excluded

Relevant academic researches show that carbon emissions during phase of operation and maintenance has major share in lifecycle followed by materials production [6]. While the accounting theory of carbon emissions during building lifecycle became mature, many studies also focus on effective carbon reduction strategies. Based on residential and commercial building case studies, Khasreen et al. (2009) summarized different LCA tools and their differences in data quality, system boundaries, inventories, impact assessments, and interpretation of results in life cycle assessment [14]. Sharma et al. (2011) reviewed the effects of different building types, life cycle stages, and locations on the environment or energy consumption [15]. Stazi et al. (2012) proposed an optimized method for integrating building energy consumption and environmental impacts on complex building skins, validating the possibility of using LCA methods to optimize carbon emissions [16]. She et al. (2014) proposed specific emissions reduction measures for each stage based on the results of estimation [17]. Zhang et al. (2014) explored the effects of life cycle energy consumption and carbon emissions on residential buildings from three categories: time difference, regional difference and technical difference [18]. Zhang et al. (2015) and Wang et al. (2015) compared the carbon emissions of buildings under different structural types [11,19]; Zhang et al. (2016) proposed a carbon reduction strategy on insulation thickness and building service life [20]; Peng (2016) and Ma et al. (2017) used sensitivity analysis methods to determine the significant influencing factors of carbon emissions [21,22]; Peng (2016) and Yang et al. (2018) also constructed the building information model (BIM) for their cases to facilitate data acquisition and calculation of carbon emissions [22,23].

However, due to limitations of data collection, and the large range of construction techniques and material choices, it is difficult to compute carbon emissions in all phases and processes during lifecycle [24]. Not to mention calculating that in the early stage of design, which is regarded as the effective time to reduce the carbon emissions of buildings [25]. When calculating the carbon emissions during building's lifecycle, usually, there are conflicts between details of system boundary and difficulties of data acquisition. For this problem, this study collected several approaches to data acquisition during different phases in relevant studies, as shown in Table 1. For built buildings with sufficient data, manufacture and operation phase are generally derived from statistical data [21,26];for uncompleted buildings or buildings with difficulties in data acquisition, data of manufacture phase generally comes from BIM, software for quantity survey or drawings [22,25,27,28], data of operation phase comes from software simulation [20,22,29,30] or reference value [28]. Generally speaking, erection and demolition phase are sometimes ignored because of their small proportion in this life cycle [25,29]. For the erection phase, the data is more from bill of quantities (BOQ) [21] or BIM [22], and some studies use Estimated or predicted value [30,31]. Demolition and end-of-life of materials, however, are seldom included in life cycle studies of buildings because lack of data [29], estimated or predicted value are used when calculation is required. Several studies have shown that using regression models based on designing factors (such as floor number, floor area) to estimate the carbon emissions during embodied or demolition phase is effective [5,32].

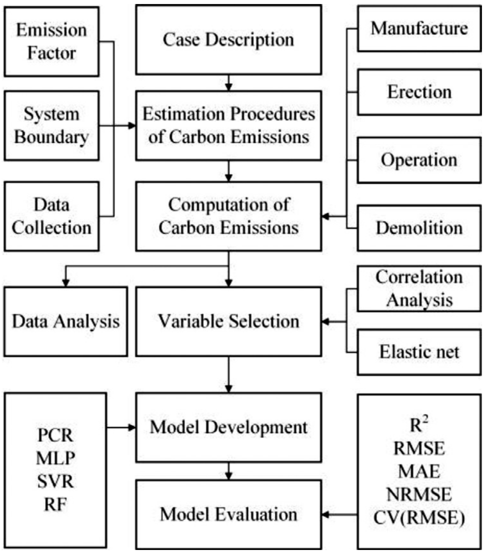

Fig. 1. The roadmap of this study.

In summary, more detailed calculation of carbon emissions generally requires reliable data of erection and operation phase from built buildings. But during stage of project design, BIM and predicted models, reference values, and simulation software are often used when carbon emissions need to be accounted. Even though these studies provide ideas to calculate carbon emissions during building's lifecycle, it is still complicated to do that at early stage of architectural design. In view of this problem, this paper attempted to give a predictive regression model of carbon emissions during building's lifecycle based on designing factors. The roadmap of this paper is organized as Fig. 1. In section 2, by the process analysis method, this paper calculated and analyzed life cycle carbon emissions of 207 residential buildings in Tianjin. Section 3 provides an introduction of regression techniques, variable selection, and evaluation indices. Section 4 shows the development of regression mod-

Table 2
Distribution of floor area and floor number of case buildings.

Floor area/m^2	Percentage/%	Floor number	Percentage/%
500–6500	57.97	3–6	8.70
6500–12,500	32.37	7–12	37.68
12,500–18,500	7.25	13–18	21.74
18,500–24,500	1.93	19–24	11.11
24,500–30,000	0.48	25–34	20.77

els respectively. In Section 5, the results of model development are analyzed and discussed. Finally, conclusions and limitations are given in Section 6.

2. Materials

2.1. Case description

Located in the eastern part of the North China Plain, Tianjin is dominated by a warm temperate semi-humid monsoon climate with four distinct seasons. In spring, the wind is dry and windy. In summer, it is mostly southerly, and it is hot and humid. The autumn is cool and pleasant. In winter, the northwest wind is prevailing, the weather is cold and dry. The annual average temperature is 12–15 °C; the coldest in January, the average temperature is −5 to −2 °C; the hottest in July, the average temperature is 26–28 °C. These parameters place Tianjin in China's cold region.

This paper collected 207 case buildings, which were built from 2013 to 2018, and the structural types are all reinforced concrete (RC). From Table 2, we can see that 58% of the case buildings are less than 6500 square meters, and they are concentrated in middle- and high-rise residential buildings, and just 8.70% are low-rise residential and multi-story residential buildings. This case study covers a wide range of floor area, building parameters and structural types. It not only helps to find characteristics of carbon emissions from residential buildings in Tianjin, but also provides a reference for the study of carbon emissions of residential buildings under similar climate.

2.2. Procedure for carbon emissions accounting

Assessment of carbon emissions during building's lifecycle covers all CO_2 equivalent emissions ($CH_4$and N_2O emissions were also considered in this paper using the relevant 100-year global warming potential [25]) at different phases of the lifecycle. The system boundary is shown in Table 3. The total carbon emissions of lifecycle can be calculated by formula (1). Specific accounting models for each stage were chosen from relevant study [28], they won't be involved in this paper.

$$E_{LifeCycle} = E_{Manufacture} + E_{Erection} + E_{Operation} + E_{Disposal} \tag{1}$$

where $E_{\mathrm{LifeCycle}}$, $E_{\mathrm{Manufacture}}$, E_{Erection}, $E_{\mathrm{Operation}}$ and E_{Disposal} are carbon emissions during lifecycle, manufacture, erection, operation and disposal, respectively, $kgCO_2/m^2$.

Process analysis method was used in inventory analysis during life cycle assessment. In this case, it is necessary to define clear boundary conditions based on carbon emission calculation and calculate all the factors in the process. At the same time, the CO_2 emission factors were used directly to calculate carbon emissions in each process (Formula (2)). CO_2 emission factors were selected or referenced from relevant studies [11,28,34], as shown in Table 4.

Table 4
Major carbon emission factors covered in this paper.

Type	Name	Unit	Value
Energy	Natural gas	$kgCO_2/m^3$	2.34
	Gasoline	$kgCO_2/kg$	3.51
	Electricity	$kgCO_2/kWh$	0.725
Transport vehicles	Gasoline truck	$kgCO_2/(10^4$ t•km)	2007.93
	Concrete(C30)	$kgCO_2/m^3$	266.22
	Steel	$kgCO_2/t$	2062.38
	Mortar(M2.5)	$kgCO_2/m^3$	214.35
Building materials	Block	$kgCO_2/m^3$	156.93
	Insulation	$kgCO_2/t$	17399.85
	Glass	$kgCO_2/t$	965.50

$$E = \mathrm{AD} \times \mathrm{EF} \tag{2}$$

where E is carbon emissions, AD is activity data, EF is emission factor.

The consumption data of materials mainly based on the architectural drawings or BOQ. However, it is rarely possible to consider all building materials in the analysis owing to the complexity and diversity of their properties, only the core and representative types of materials would be considered [35]. Based on statistical data from several case buildings, six materials such as concrete, steel, mortar, block, insulation and glass were considered at last, which could account for about 90% of total carbon emissions from all building materials in BOQ.

The erection phase includes the transportation of materials, mechanical use and energy consumption in construction processes. For the convenience of calculation, the transportation mode was chosen as gasoline trucks uniformly; based on location analysis of major factories of building materials in Tianjin, the transportation distance was uniformly set as 30 km. The data of machinery and equipment application was taken from the BOQ, and for the case that lacking BOQ, the data was estimated by building levels as follows [32]:

$$E_{Erection} = X + 1.99 \tag{3}$$

where X is building levels aboveground.

The operation phase includes building's operation and replacement of materials. The lifespan was assumed to be 50 years, and DesignBuilder was used to simulate energy consumption during building's operation. The parameters in building performance simulation were set up uniformly, as shown in Table 5 [36,37]; as for HVAC, the cooling load is borne by the room air conditioner, and the fuel for heating boiler was selected as natural gas. Because other selected materials have a lifespan more than 50 years [17], only the carbon emissions generated by replacing glass were considered.

The disposal phase includes the process of demolition, waste transportation and disposal. Among them, energy consumption of demolition is usually ignored [21,25,29], and this paper used carbon emission factor of 2.52 kg CO_2/m^2 (floor area) for calculation

Table 3
System boundary.

Phases	Details
Manufacture	The manufacture of building materials
Erection	Transportation of building materials, machinery & equipment application
Operation	Energy consumed during building's operation, replacement of materials
Disposal	Demolition, waste transportation, waste recycling

Table 5
Parameters setting in building performance simulation.

Design condition		Load density			HVAC	
Summer/°C	Winter/°C	Occupancy	Lighting	Equipment	Cooling	Heating
26	18	40 m^2/person	6 W/m^2	5 W/m^2	Refrigeration COP = =3.00	Boiler efficiency = 0.80

[17]; calculation of waste transportation is similar to the transportation of materials, and the total weight of waste was taken as 80% of the total weight of materials [38]; only the recycle of steel was considered, the carbon emission factor of steel reprocessing is 578.36 $kgCO_2$/t [32].

2.3. Computation of carbon emissions

In this paper, the annual carbon emissions per unit of floor area "$kgCO_2$/(m^2•year)" was used as the accounting unit. According to the results, carbon emissions from case buildings' lifecycle ($LCCO_2$) are 30.0–60.0 $kgCO_2$/(m^2•year), carbon emissions during each phase are shown in Fig. 2. Carbon emissions during operation phase account for the largest proportion (75%–87%) in the lifecycle, and the manufacture phase is second (11%–25%). As two phases with less carbon emissions, erection phase and disposal phase account for 1.14%–6.65% and −1.44% to 0.96%, respectively. The results are similar to other studies in China [11,20,21,27,28,39]. Fig. 3 shows the results of carbon emissions assessment for six residential building in other case studies. Due to the difference of phase division between different studies, the lifecycle was divided into the materialization phase (including material production, transportation and construction), the operation phase and the disposal phase. The $LCCO_2$ in these cases ranged from 20 to 80 $kgCO_2$/(m^2•year), with the operation phase accounting for 58.14%–83.85% of the lifecycle and the disposal phase being −0.37% to 3.23%.

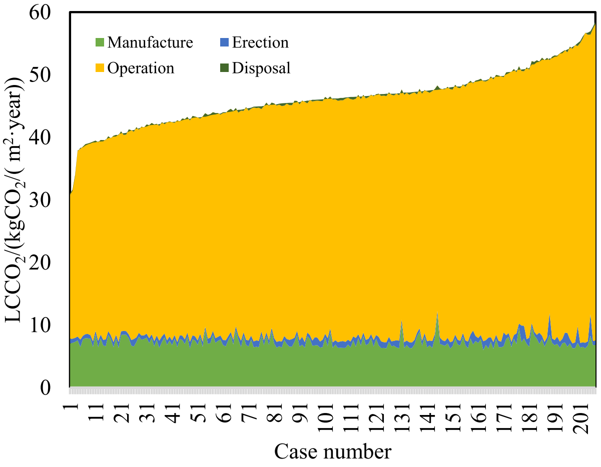

Fig. 2. Carbon emissions from different stages of building lifecycle.

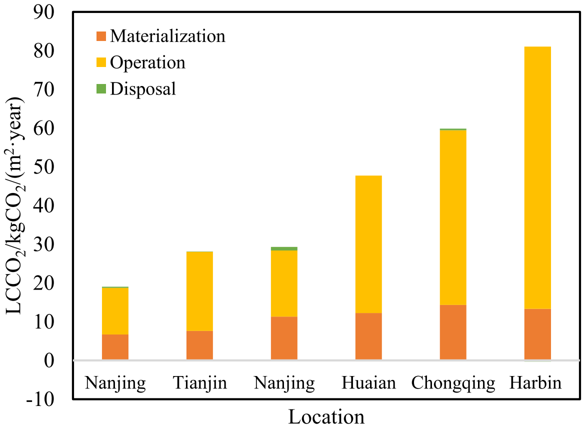

Fig. 3. Carbon emissions from building lifecycle in other studies.

3. Methodology

In this section, four selected regression techniques were introduced firstly; methods of variable selection were then introduced for improving predictive accuracy and interpretability of regression models; at last, several evaluation indices were introduced for evaluating the accuracy of each model.

3.1. Regression techniques

3.1.1. PCR

Principal Component Analysis (PCA) was first proposed by Karl Pearson [40], that's a dimensionality reduction algorithm, which converts high-dimensional data into low-dimensional data with minimal loss. PCA eliminates the multicollinearity between variables by orthogonal transformation and forms a new variable, principal component. These principal components retain the vast majority of the original variables and have no correlation with each other. Therefore, regression model can be established by these "new variables" based on the least square method, which is called principal component regression (PCR) [41].

3.1.2. MLP

Multi-Layer Perceptron (MLP) is a kind of feed-forward artificial neural network, a supervised learning technique called Back-Propagation (BP) is usually utilized to train MLP. Neural network originates from the study of nervous system, they are composed of interconnected neurons (also called processing elements), which are usually arranged in three layers (or more); the input layer and the hidden layer receive the input vector, and the output layer produces the output, as shown in Fig. 4.

BP model contains two important processes: (1) forward the input signal to the output layer through a nonlinear activation function; and (2) adjust the connection weights by transmitting predictive errors in the reverse direction. These two processes will be performed in a cyclic alternating mode until the predictive error converges to a stable value. The information function transmitted from the ith layer to the jth layer is expressed as follow [42]:

$$y_j^m(n) = f\left(\sum_{k=1}^{p} w_{ji}^k(n) y_i^k(n)\right) \tag{4}$$

where, $y_j^m(n)$ is the output value of the m-th neuron of the jth layer; $f(x)$ is the activation function; p is the total number of nodes in the ith layer; $w_{ji}^k(n)$ is the connection weight from the k-th node of the ith layer in the n-th iteration.

The function of back-propagation process is expressed as:

$$w(n+1) = w(n) - \eta \frac{\partial E}{\partial w(n)} + \alpha \Delta w(n) \tag{5}$$

where, $w(n)$ is the weight at the nth iteration; η is the learning rate; E is the predictive error energy; α is the momentum parameter that can accelerate the convergence.

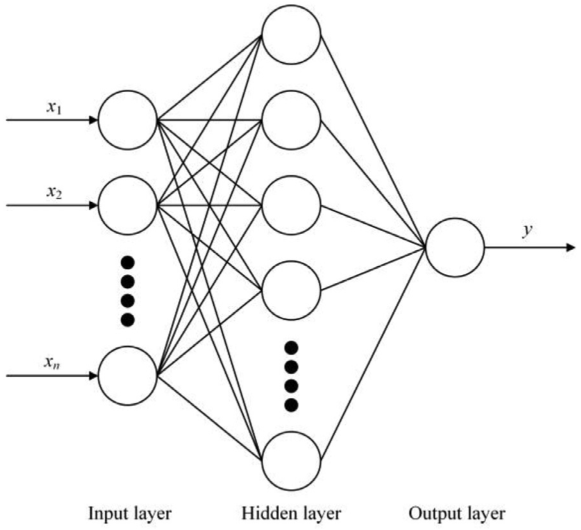

Fig. 4. Structure of multi-layer perceptron.

3.1.3. SVR

Support vector machine (SVM) is a learning algorithm proposed in 1992 [43]. It is widely used to solve classification problems, especially for problems of small sample, high dimension and nonlinear data [42]. Support vector regression (SVR) can be used to study the regression relationship between independent variables and continuous dependent variable, and solve forecasting problems.

The objective function of SVR can be defined as follow [44]:

$$\begin{aligned} &\min_{w,b} \tfrac{1}{2}\|w\|^2 + \tfrac{C}{n}\sum_{i=1}^{n}\left(\xi_i^{\,2} + \xi_i^{*2}\right) \\ &s.t.(b + W^T X_i) - y_i \le \varepsilon + \xi_i, i = 1, 2, ..., n \\ &y_i - (b + W^T X_i) \le \varepsilon + \xi_i^*, i = 1, 2, ..., n \\ &\xi_i \ge 0, \xi_i^* \ge 0, i = 1, 2, ..., n \end{aligned} \tag{6}$$

where, the regularization parameter C and the slack variables ξ_i and ξ_i^* are introduced. ξ_i and ξ_i^* are the slack variables above and below the sample observation point, respectively, defined as:$\xi_i = \max(0, |e_i| - \varepsilon)$. Moreover, in order to reduce the risk of overfitting, the ε-insensitive loss function is also introduced.

If the Lagrangian multiplier is introduced to convert the above formula into a dual problem, it will be easier to solve. The corresponding regression function can be written as follow [45]:

$$f(x) = \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)K(x_n, x) + b \tag{7}$$

where, α_n and α_n^* are non-negative multipliers; x_n is sample observation point; $\sum_{n=1}^{N}(\alpha_n - \alpha_n^*)$=0, $K(x_n, x)$ is the kernel function.

3.1.4. RF

Random Forest (RF) is a commonly used machine learning algorithm proposed by Leo Breiman in 2001 [46]. It is flexible and practical, and good at handling high-dimensional data. RF is based on the idea of ensemble learning to integrate many decision trees, which belongs to the ensemble learning method.

RF integrates all the classification voting results, and specifies the category with the highest number of votes (classified target variable) or the average value (continuous target variable) as the final output. The essence of RF is to apply the bootstrap method (replacement sampling method) to the CART (Classification and Regression Tree) algorithm, that is, the original training sample is sampled back, so that the sample size is as same as the original one. The training dataset uses the undrawn samples as an Out-Of-Bag (OOB) Dataset for evaluation [47]. In sampling, not only the cases in the original data set are sampled, but also the set of all p independent variables are sampled by random sampling, and then a subset of all input variables ($X_{1,}$ X_2,..., X_m) is obtained., ($m \ll p$, generally $m = \sqrt{p}$ or $m = p/3$). Then the CART algorithm is used to construct the decision tree model for the m independent variables of the obtained bootstrap samples. All trees in the forest will grow as much as possible and are not considered for pruning. After repeating this steps times, all the resulting decision tree models are integrated together, which is called RF model [47].

3.2. Variable selection

Correlation analysis can analyze two or more related variables, and tend to measure the closeness of two variables. However, when dealing with high-dimensional data, it often encounters problems such as multi-collinearity between variables and high noise. These problems easily lead to over-fitting of regression models, which will reduce the predictive accuracy and interpretability of the model. So more effective method of variable selection is required before processing high-dimensional data. The contraction method is often used to select variables, it mainly includes ridge regression, least absolute and selection operator (LASSO), and elastic net [48].

Elastic net algorithm, first proposed by Zhou and Hastie in 2005, was utilized in this paper [49]. This algorithm can achieve both variable selection and cluster selection, its essence is to combine ridge regression and LASSO into a single model with two penalty factors: the L_1 norm and the L_2 norm are constrained on the basis of least square. The function is expressed as follows:

$$\hat{\beta}_{Enet} = \arg\min_{\beta}\left\{\left\| y - \sum_{i=1}^{p} X_i\beta_i \right\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2\right\} \tag{8}$$

where λ_1 and λ_2 are non-negative and adjustable parameters, and can also be written as:

$$\begin{aligned} &\hat{\beta}_{Enet} = \arg\min_{\beta}\left\{\left\| y - \sum_{i=1}^{p} X_i\beta_i \right\|^2\right\} \\ &s.t.\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2 \le t \end{aligned} \tag{9}$$

Elastic net can not only perform continuous compression and automatic variable selection like LASSO, but also process parameter estimates of correlation between variables like ridge regression [50]. In general, elastic net can achieve higher predictive performance, but due to the introduction of more control parameters, the workload of determining parameter values based on data is actually larger [51].

3.3. Evaluation indices

In order to evaluate the accuracy of each regression model, several evaluation indices were considered, as shown in Table 6 [52].

Table 6
Evaluation indices of predictive accuracy for regression model.

Evaluation indices	Abbreviation
Coefficient of determination	R^2
Root mean square error	RMSE
Mean absolute error	MAE
Normalized root mean square error	NRMSE
Coefficient of variance of the root mean square error	CV(RMSE)

The function of each evaluation indice is expressed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{10}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \tag{11}$$

$$MAE = \frac{\sum_{i=1}^{n} \left|\hat{y}_i - y_i\right|}{n} \tag{12}$$

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}}{y_{\max} - y_{\min}} \tag{13}$$

$$CV(RMSE) = \frac{\sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}}{\bar{y}} \tag{14}$$

where, $\hat{y}_i$ and y_i are predictive and actual value; $\bar{y}$ is average value; $y_{\max}$ and $y_{\min}$ are the maximum and minimum value, n is the number of samples.

4. Model development

4.1. Variable selection

In order to consider enough variables to explain the dependent variable ($LCCO_2$), this paper pre-selected 17 design variables based on related studies [53–56], as shown in Table 7. The analysis found that there are outliers in the data (that is, there are unreasonable values in the dataset). Outliers may adversely affect data analysis and model development, such as increasing error differences and reducing the ability of statistical tests. Therefore, some outliers were eliminated in this paper.

Due to 17 selected variables are not necessarily all valuable, we need to choose explanatory variables that have a significant impact on the dependent variable to avoid redundant information and noise interference. Firstly, this paper measured the degree of correlation between independent variables and dependent variable, as shown in Fig. 5.

Table 8
Coefficient value of elastic net at step 422.

Variable	Coefficient	Variable	Coefficient
X_1	−0.006	X_{10}	0.116
X_2	0	X_{11}	−0.007
X_3	0	X_{12}	0
X_5	−0.035	X_{13}	0.005
X_6	−0.058	X_{14}	0.098
X_7	−0.086	X_{15}	0.096
X_8	0	X_{16}	0.194
X_9	0.033	X_{17}	0.192

It can be seen from Fig. 5 that the correlation between $LCCO_2$ and independent variables except X_4 (standard floor area) is significant; Moreover, it can be found that there is a very serious collinearity between independent variables. The existence of multi-collinearity tends to lead to inaccuracies in the evaluation model and cannot reflect the true correspondence.

So, elastic net was chosen for further variable selection, it was used to analyze the normalized data, and the optimal solution is obtained at step 422. The regression coefficients are shown in Table 8. According to the results, 12 variables of X_1, X_5, X_6, X_7, X_9, X_{10}, X_{11}, X_{13}, X_{14}, X_{15}, X_{16}, and X_{17} were selected as independent variables for the regression models.

In the following content, these 12 independent variables will be used to development predictive model by PCR, MLP, SVR and RF respectively, as shown in Fig. 6.

4.2. PCR model development

PCR is one of the common methods for dimensionality reduction, which can effectively solve the multi-collinearity problem. First, PCA was performed on 12 independent variables, and the percentage of variance and principal components were as shown in Tables 9 and 10. The results show that the method extracted four principal components, and the cumulative variance contribution is 86.17%. The loads of the four principal components on each variable are different, and the actual meaning is not very clear: The coefficients of floor number, building height, floor area, building volume, etc. in the principal component F_1 are larger, so it may be reflect the comprehensive index of building volume; The coefficients of the heat transfer coefficient of roof, external wall and glass in the principal component F_2 are larger, which may reflect energy conservation performance of building envelope; In the principal component F_3, the coefficients of shape coefficient and body

Table 7
Variables interpretation.

Variable name	Implication	Abbreviation	Unit
X_1	Floor number	N(Story)	–
X_2	Story height	H(Story)	m
X_3	Household number	N(Household)	–
X_4	Area of typical floor	S	m^2
X_5	Building height	H	m
X_6	Floor area	A	m^2
X_7	Building volume	V	m^3
X_8	Building envelope surface area	S(Surface)	m^2
X_9	Shape coefficient, building envelope surface area/ building volume	SC	m^{-1}
X_{10}	Body coefficient, building envelope surface area/ construction area	BC	1
X_{11}	Building height/ Standard floor area	H/S	m^{-1}
X_{12}	South-facing window-to-wall ratio	WWR(S)	–
X_{13}	North-facing window-to-wall ratio	WWR(N)	–
X_{14}	West & East-facing window-to-wall ratio	WWR(W&E)	–
X_{15}	Heat transfer coefficient of roof	K(Roof)	$W/(m^2\ K)$
X_{16}	Heat transfer coefficient of external wall	K(Wall)	$W/(m^2\ K)$
X_{17}	Heat transfer coefficient of glass	K(Glass)	$W/(m^2\ K)$
Y	Carbon emissions during building's life-cycle	$LCCO_2$	$kgCO_2/(m^2\ year)$

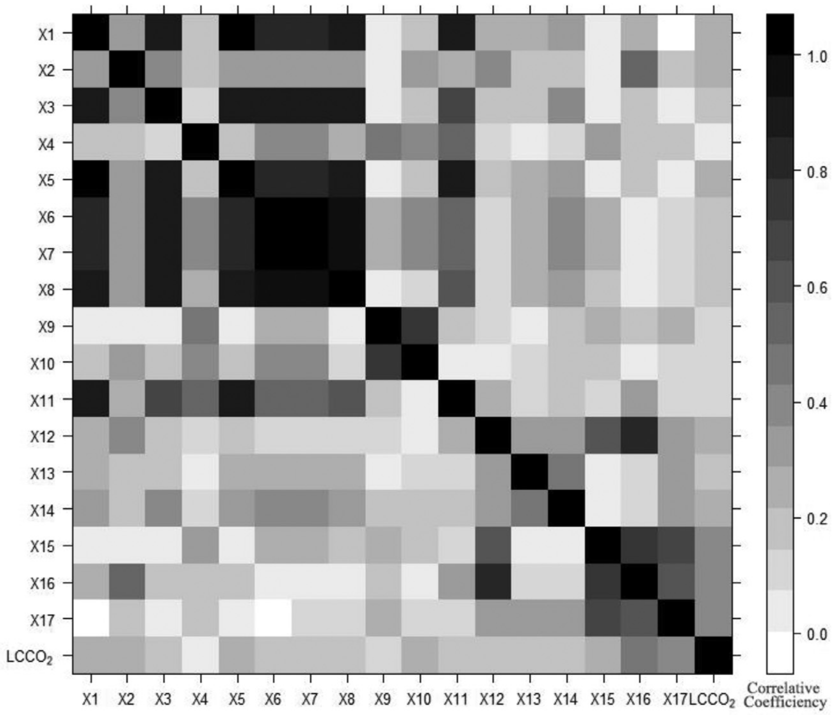

Fig. 5. Correlation analysis between variables.

Table 9
Total variance explained.

Component	Initial eigenvalues			Extraction sums of squared loadings		
	Total	Percentage of variance %	Cumulative %	Total	Percentage of variance %	Cumulative %
1	4.644	38.696	38.696	4.644	38.696	38.696
2	2.791	23.262	61.958	2.791	23.262	61.958
3	1.575	13.124	75.083	1.575	13.124	75.083
4	1.331	11.088	86.170	1.331	11.088	86.170
5	0.557	4.641	90.811			
6	0.487	4.056	94.868			
7	0.242	2.019	96.886			
8	0.195	1.626	98.512			
9	0.169	1.409	99.921			
10	0.008	.064	99.985			
11	0.002	.015	100.000			
12	2.493E−05	.000	100.000			

Table 10
Component matrix.

Variable	Component				Variable	Component			
	1	2	3	4		1	2	3	4
N(Story)	0.950	−0.220	0.116	−0.144	H/S	0.755	−0.422	0.254	−0.129
H	0.946	−0.223	0.125	−0.153	WWR(N)	0.374	0.150	0.110	−.753
A	0.923	0.127	−0.071	−0.121	WWR(W-E)	0.484	0.206	0.008	0.669
V	0.921	0.127	−0.065	−0.130	K(Roof)	0.122	0.824	0.316	−0.273
SC	−0.175	−0.611	−.675	0.028	K(Wall)	−0.153	0.753	0.434	−0.300
BC	−0.336	−0.442	−.744	0.096	K(Glass)	0.137	0.777	0.404	0.221

coefficient are larger, which could reflect the influence of building shape; The coefficients of the north-facing and the east- and west-facing window-wall ratio of the principal component F_4 are larger, which reflect the factor of building window-wall ratio.

Then, the extracted four principal components were used as "new variables" instead of the original variables for multiple linear regression analysis, and the regression equation was obtained as follows:

$$LCCO_2 = 46.331 - 0.806 \times F_1 + 1.340 \times F_2 + 1.761 \times F_3 + 0.996 \times F_4 (R^2 = 0.352) \quad (15)$$

4.3. MLP model development

In fact, the use of MLP in this paper has certain limitations. First, MLP is suitable for the case that data sample size is sufficient (at least 10 times more than the number of variables included in the model [47]), and too many variables may lead to over-fitting problems; secondly, the determination of the number of hidden layer nodes in the neural network is still inconclusive and can be determined gradually in the analysis.

The parameters were set as follows: (1) The partition has training set, validation set, and test set, the ratio was generally taken as

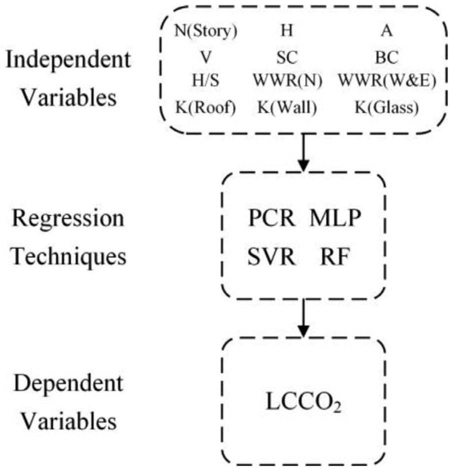

Fig. 6. Model development.

6:2:2; (2) Set the number of hidden layers to single layer and double layer for comparison analysis, where 4 groups were set in the single layer and 6 groups were set in the double layer, as shown in Fig. 7. (3) On the basis of comparison analysis, the initial learning rate was set to 0.1, the boundary minimum was 0.001, and the momentum parameter was set to 0.95; (4) the activation function was set to hyperbolic tangent function. The test set was used to measure the predictive performance of model. Then, the relative error (the squared sum of the residuals) of the test set in the 10 pre-trainings was set in the different numbers of hidden layers and nodes, the results were recorded in Figure. 11. When the number of hidden layers was 2 and the number of nodes was 20 + +15, the model performed best.

4.4. SVR model development

In this paper, the radial basis function (RBF) was used to train SVR model. At the same time, the regularization constant C and the kernel function parameter γ need to be set reasonably. C is loss penalty term; if C is too large, the convergence of SVR would be difficult and unstable, if the value of C is too small, it will take more time to converge. Increasing the parameter γ can improve the predictive accuracy, but may lead to over-fitting; decreasing the parameter γ will reduce the deviation, but will cause the instability. The training data were used for the parameter selection. The search for optimal C and γ ranged from $[2^{-6}, 2^{6}]$ with 2^{1} as the exponential step size growth [57]. Consequently, a total of 196 different combinations of {C, γ} were generated and tested. The mean square error in the 10-fold cross-validation results was used as evaluation to select the optimal {C, γ}. In this research, the optimal C and γ were selected as 4 and 0.0625.

4.5. RF model development

The RF model development requires three user-defined parameters to be determined. They are: the minimum number of terminal nodes for each tree (nodesize), the number of trees in the forest (ntree), and the number of randomly selected variables to grow the tree (mtry) [58].

The nodesize parameter controls the size of each tree within the forest. Essentially, the selection of this parameter determines when to stop the tree splitting process. A large nodesize will cause shallow trees because of limits in the tree splitting process, the predictive accuracy of each tree could not be guaranteed. On the contrary, a small nodesize would bring deep tree structure which creates comprehensive learning from the data, which would cost more computation time and may encounter overfitting [57,58]. This study used the default value of 5, which is also a commonly used and recommended value [47].

The ntree parameter determines the number of trees generated in an RF model. A larger ntree will improve the predictive performance of RF because more trees would be considered and the problem of overfitting would be avoided, but the calculation time would be increased at the same time. In this paper, ntree is set to 800. It was found that when the number of trees is more than 500, the mean square error of regression model has become stable. Mtry affects the accuracy of predictive by introducing randomness in the model development, which affects the predictive accuracy of the decision tree in the forest and the correlation between decision trees. Usually, the larger the mtry is, the better the predictive performance will be, but that will increase the correlation between decision trees. In this paper, mtry was set to 1–12, then the model's interpretable variance percentage and residual mean square were obtained, as shown in Fig. 8. The regression model has optimal predictive accuracy when mtry was set to 7.

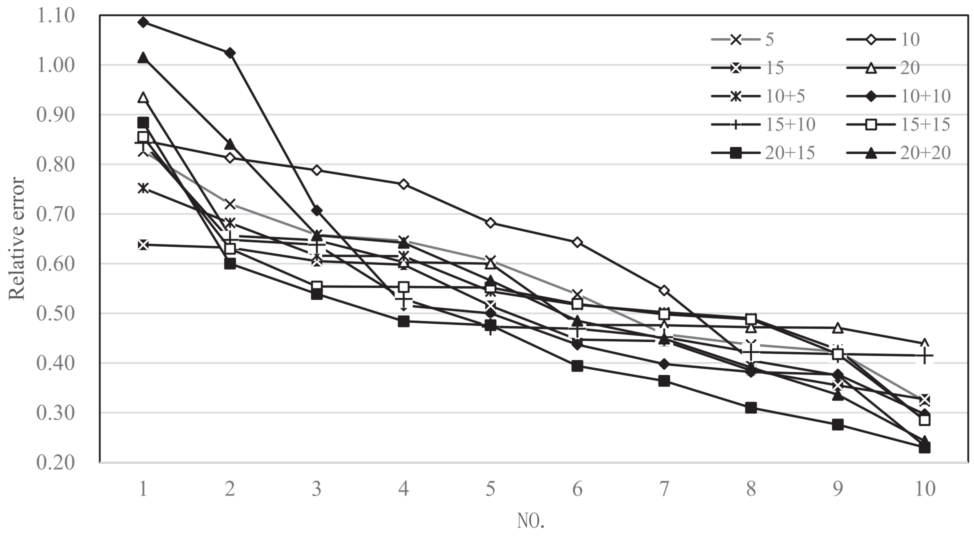

Fig. 7. Relative error of each test set under different layers and number of nodes.

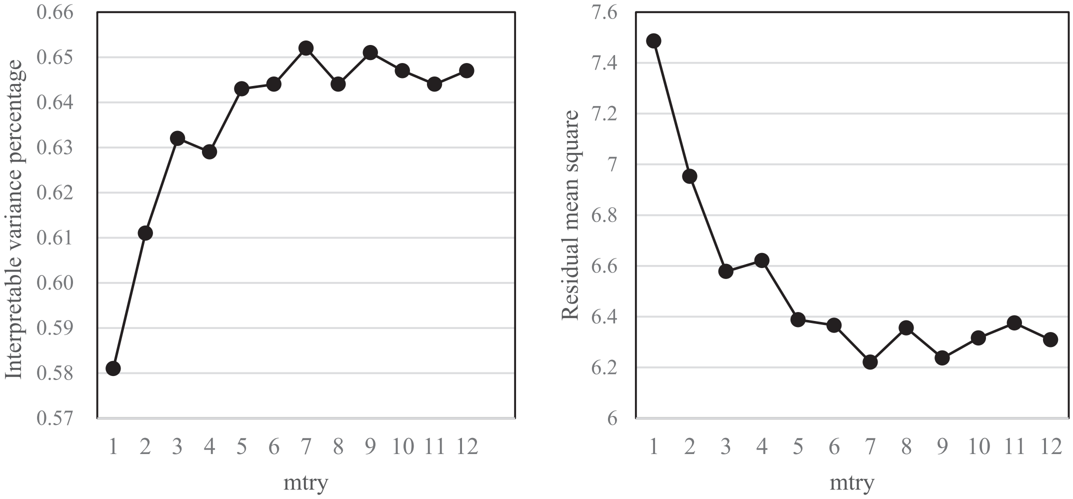

Fig. 8. Interpretable variance percentage and residual mean square under different mtry.

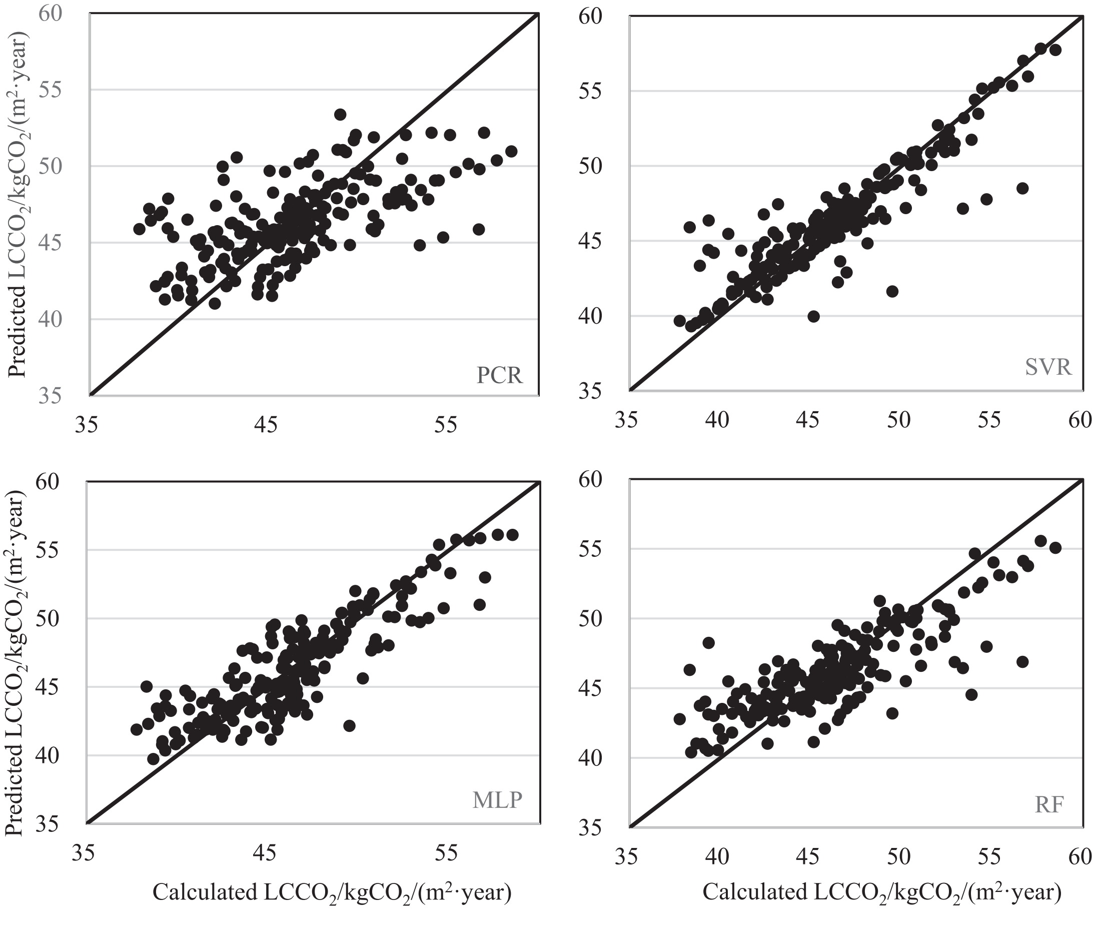

Fig. 9. Comparison of predicted and calculated values of LCCO$_2$

5. Results and discussion

In this paper, PCR, RF, MLP and SVR were used to develop regression model of LCCO$_2$ respectively. After process analysis and parameter setting previously, the predictive results were obtained. Comparison of the predicted values from each model and calculated values of LCCO$_2$ is shown in Fig. 9. In order to visually see the difference of predictive performance, 25 case buildings were selected randomly to compare together, as shown in Fig. 10. From the above two pics, the degree of deviation between the predicted value and calculated value of each regression model can be seen. The evaluation indices of each regression model are shown in Table 11.

From the process of model development and analysis, the following conclusions can be drawn:

1) Predictive performance. According to the evaluation indices, the predictive performance of regression models can be ranked as: SVR>MLP>RF>PCR, and SVR performed best, which verifies its performance when dealing with problems with small samples and high dimensions. Secondly, MLP shows that the well-trained or -learned neural network model also perform well;

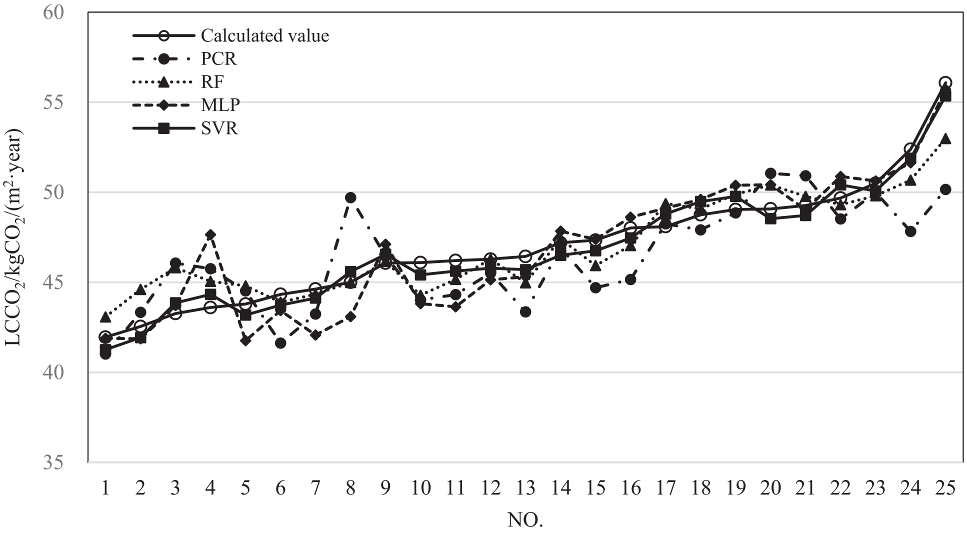

Fig. 10. Comparison of calculated values and predicted values of 25 case buildings.

Table 11
Evaluation indices of regression models.

Model	R^2	RMSE/($kgCO_2/(m^2$ year))	MAE/($kgCO_2/(m^2$ year))	NRMSE	CV(RMSE)
PCR	0.365	3.368	2.541	0.163	0.073
RF	0.651	2.494	1.823	0.120	0.054
MLP	0.744	2.136	1.648	0.103	0.046
SVR	0.800	1.889	1.206	0.091	0.041

the predictive performance of RF is slightly lower than the former two. The reason may be the discretization of continuous variables in the decision tree generation leads to a decrease in the number of nodes, so that the newly generated data set loses part of original information [59]; the PCR model has the worst performance.

2) Multi-collinearity. Garg et al. indicated that the machine learning technique is suitable for tackling the multi-collinearity problem [60]. The results show that machine learning techniques such as RF, MLP and SVR are good at solving multi-collinearity problems and all perform better than PCR.
3) Process of development. From the perspective of development process, PCR is the most stable and fastest, but its predictive performance is lowest. The development of RF, MLP and SVR models is slightly more complicated, which not only requires longer calculation time and complicated parameter adjustment, but also the difference in stability of the model's generalization ability. The RF model is very stable in development, and as the number of ntree increases, the model tends to be stable and over-fitting can be avoided. MLP is based on a gradient descent algorithm and therefore easily limited to local optima, particle swarm optimization (PSO) algorithm and genetic algorithm (GA) are often used to optimize this problem. The performance of SVR is between the above two, it can achieve better predictive performance after reasonable setting of parameters, and performed best in this study.

6. Conclusion

Conventional calculation methods of carbon emissions during building's lifecycle rely on detailed data of various phases, the specific data from erection and operation phase would make the results more reliable. But in this case, it will be inconvenient and complicated to estimate carbon emissions during lifecycle at the beginning of architectural design. To solve this gap, this paper attempted to develop regression model of carbon emissions during building's lifecycle using designing factors. Based on computation of carbon emissions from case residential buildings, four types of models (PCR, MLP, SVR, and RF) were utilized to develop models, these techniques are capable of mining the complex relationship among $LCCO_2$ and selected designing factors with the aid of their nonlinear mapping capacity, self-learning ability and generalization ability. Then a comprehensive comparative study was conducted to test the performance of these models. The rank of predictive accuracy is SVR>MLP>RF>PCR. The proposed model provides the possibility of a simple and easy estimation of carbon emissions and suitable for the early design phase.

However, certain limitations of this study have to be mentioned. First, some simplifications and assumptions were made in the calculations of $E_{LifeCycle}$ due to the lack of relevant statistical data. This should be improved along with the enrichment of data source. Second, the performance of these models depends significantly on the parameters setting, this study used traditional optimization methods for parameter optimization, intelligent optimization approaches, such as PSO and GA, should be explored in future works. Limited by ability and time, this paper only conducted research on residential buildings in Tianjin. In the follow-up work, the characteristics of carbon emissions and influencing factors of different regions and building types should be studied.

Declaration of Competing Interest

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

Acknowledgment

This research was funded by the National Key Research and Development Program of China (Grant No. 2016YFC0700201) and Science and Technology Research Programs of the Hebei Institutions of Higher Education (QN2019187).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.enbuild.2019.109519.

References

[1] The Intergovernmental Panel on Climate Change, Global warming of 1.5 °C – summary for Policymakers, 2018 [Online]. Available: https://www.ipcc.ch/sr15/. [Accessed 15 November 2018].

[2] J.H. Yuan, Y. Xu, X.P. Zhang, Z. Hu, M. Xu, China's 2020 clean energy target: consistency, pathways and policy implications, Energy Policy 65 (2014) 692–700.

[3] United Nations Environment Programme, Buildings and climate change: summary for decision-makers, 2009 [Online]. Available: https://europa.eu/capacity4dev/unep/document/buildings-and-climate-change-summary-decision-makers. [Accessed 15 November 2018].

[4] Building Energy Research Center At Tsinghua University, China Building Energy Efficiency Annual Report on 2017, China Building Industry Press, Beijing, 2017.

[5] Z.X. Luo, L. Yang, J.P. Liu, Embodied carbon emissions of office building: a case study of China's 78 office buildings, Build. Environ. 95 (2016) 365–371.

[6] T. Ramesh, R. Prakash, K.K. Shukla, Life cycle energy analysis of buildings: an overview, Energy Build. 42 (2010) 1592–1600.

[7] M. Suzuki, T. Oka, Estimation of life cycle energy consumption and CO_2 emission of office buildings in Japan, Energy Build. 28 (1998) 33–41.

[8] S. Xing, Z. Xu, G. Jun, Inventory analysis of LCA on steel- and concrete-construction office buildings, Energy Build. 40 (2008) 1188–1193.

[9] H.J. Wu, Z.W. Yuan, L. Zhang, J. Bi, Life cycle energy consumption and CO_2 emission of an office building in China, Int. J. Life Cycle Assess. 17 (2012) 105–118.

[10] A. Stephan, R.H. Crawford, K. de Myttenaere, A comprehensive assessment of the life cycle energy demand of passive houses, Appl. Energy 112 (2013) 23–34.

[11] X.C. Zhang, F.L. Wang, Life-cycle assessment and control measures for carbon emissions of typical buildings in China, Build. Environ. 86 (2015) 89–97.

[12] M.Y. Han, G.Q. Chen, L. Shao, J.S. Li, A. Alsaedi, B. Ahmad, S. Guo, M.M. Jiang, X. Ji, Embodied energy consumption of building construction engineering: case study in E-town, Beijing, Energy Build. 64 (2013) 62–72.

[13] J. Hong, G.Q. Shen, Y. Feng, W.S. Lau, C. Mao, Greenhouse gas emissions during the construction phase of a building: a case study in China, J. Clean. Prod. 103 (2015) 249–259.

[14] M.M. Khasreen, G.P.F. Banfill, F.G. Menzies, Life-cycle assessment and the environmental impact of buildings: a review, Sustainability 1 (3) (2009) 674–701.

[15] A. Sharma, A. Saxena, M. Sethi, V. Shree, Varun, Life cycle assessment of buildings: a review, Renew. Sustain. Energy Rev. 15 (2011) 871–875.

[16] F. Stazi, A. Mastrucci, P. Munafò, Life cycle assessment approach for the optimization of sustainable building envelopes: an application on solar wall systems, Build. Environ. 58 (2012) 278–288.

[17] J.Q. She, Study on Carbon Emission and Emission Reduction Strategies of Public Buildings in Hot Summer and Warm Winter Area Based on LCA: Take Xiamen City as an Example, Huaqiao University, Xiamen, 2014.

[18] C.H. Zhang, B.R. Lin, B. Peng, Study on impact factors for life-cycle energy consumption and CO_2 emission of chinese housing in cold climate zone, Build. Sci. 30 (10) (2014) 76–83.

[19] Y. Wang, H. Zhang, L. Dong, Different structure types (heavy, light structure) and structural material in whole life cycle of building carbon emissions, Arch. Cult. (2) (2015) 110–111.

[20] Y. Zhang, X. Zheng, H. Zhang, G. Chen, X. Wang, Carbon emission analysis of a residential building in China through life cycle assessment, Front. Environ. Sci. Eng. 10 (2016) 150–158.

[21] J.J. Ma, G. Du, Z.K. Zhang, P.X. Wang, B.C. Xie, Life cycle analysis of energy consumption and CO_2 emissions from a typical large office building in Tianjin, China, Build. Environ. 117 (2017) 36–48.

[22] C.H. Peng, Calculation of a building's life cycle carbon emissions based on Ecotect and building information modeling, J. Clean. Prod. 112 (2016) 453–465.

[23] X. Yang, M. Hu, J. Wu, B. Zhao, Building-information-modeling enabled life cycle assessment, a case study on carbon footprint accounting for a residential building in China, J. Clean. Prod. 183 (2018) 729–743.

[24] Y. Ju, Y. C., Research on the building carbon emission calculation method in compliance with the theory of full lifecycle – based upon statistical analysis of CNKI's domestic literature dated between 1997 and 2013, Hous. Sci. 34 (2014) 32–37.

[25] J. Basbagill, F. Flager, M. Lepech, M. Fischer, Application of life-cycle assessment to early stage building design for reduced embodied environmental impacts, Build. Environ. 60 (2013) 81–92.

[26] B. Rosselló-Batle, A. Moià, A. Cladera, V. Martínez, Energy use, CO_2 emissions and waste throughout the life cycle of a sample of hotels in the Balearic Islands, Energy Build. 42 (2010) 547–558.

[27] X.X. Ou, D.Z. Li, Q.M. Li, A BIM-based estimator for carbon emissions of a building at design stage, ICCREM 2017: Project Management and Construction Technology, Guangzhou, 2017.

[28] D.Z. Li, P. Cui, Y.J. Lu, Development of an automated estimator of life-cycle carbon emissions for residential buildings: a case study in Nanjing, China, Habitat Int. 57 (2016) 154–163.

[29] M. Wallhagen, M. Glaumann, T. Malmqvist, Basic building life cycle calculations to decrease contribution to climate change – case study on an office building in Sweden, Build. Environ. 46 (2011) 1863–1871.

[30] L. Gustavsson, A. Joelsson, R. Sathre, Life cycle primary energy use and carbon emission of an eight-storey wood-framed apartment building, Energy Build. 42 (2010) 230–242.

[31] X. Wu, B. Peng, B. Lin, A dynamic life cycle carbon emission assessment on green and non-green buildings in China, Energy Build. 149 (2017) 272–281.

[32] Y.S. Zhang, Life Cycle Assessment on the Reduction of Carbon Dioxide Emission of Buildings, National Cheng Kung University, Taiwan, 2003.

[33] Y. Zhang, X. Zheng, H. Zhang, G. Chen, X. Wang, Carbon emission analysis of a residential building in China through life cycle assessment, Front. Environ. Sci. Eng. 10 (2016) 150–158.

[34] The Intergovernmental Panel on Climate Change, IPCC Guidelines for National Greenhouse Gas Inventories: vol. 2-Energy, Intergovernmenta Panel On Climate Change, 2006 [Online]. Available: https://www.ipcc-nggip.iges.or.jp/public/2006gl/. [Accessed 15 November 2018].

[35] D.W. Yu, H.W. Tan, Y.J. Ruan, A future bamboo-structure residential building prototype in China: life cycle assessment of energy use and carbon emission, Energy Build. 43 (2011) 2638–2646.

[36] Tianjin Urban-Rural Construction and Traffic Committee, Tianjin energy efficiency design standard for residential buildings (DB29-1-2013), Tianjin, 2013.

[37] Z.Q. Lu, Research on the Influencing Factors of Energy Consumption for Tianjin Residential Buildings, Tianjin University, Tianjin, 2013.

[38] Y.S. Wang, X. Yang, H. Yan, Y. Zhang, J.F. Li, Carbon emission accounting for buildings based on whole life cycle: a case study of reconstruction project at college in Guangzhou, J. Eng. Manag. 31 (3) (2017) 19–24.

[39] D.Z. Li, H.X. Chen, E.C.M. Hui, J.B. Zhang, Q.M. Li, A methodology for estimating the life-cycle carbon efficiency of a residential building, Build. Environ. 59 (2013) 448–455.

[40] K. Pearson, LIII. On lines and planes of closest fit to systems of points in space, Lond. Edinb. Dublin Philos. Mag. J. Sci. 2 (1901) 559–572.

[41] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, China Machine Press, Beijing, 2018.

[42] H. You, Z. Ma, Y. Tang, Y. Wang, J. Yan, M. Ni, K. Cen, Q. Huang, Comparison of ANN (MLP), ANFIS, SVM, and RF models for the online classification of heating value of burning municipal solid waste in circulating fluidized bed incinerators, Waste Manag. 68 (2017) 186.

[43] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, ACM, 1992, pp. 144–152.

[44] M.O. Efe, A comparison of ANFIS, MLP and SVM in identification of chemical processes, in: 2009 IEEE Control Applications (CCA) & Intelligent Control (ISIC), St. Petersburg, 2009, pp. 689–694.

[45] Understanding support vector machine regression [Online]. Available: https://ww2.mathworks.cn/help/stats/understanding-support-vector-machine-regression.html#buytaw5. [Accessed 15 November 2018].

[46] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[47] W.T. Zhang, W. Dong, SPSS Statistical Analysis Advanced Tutorial, Higher Education Press, Beijing, 2018.

[48] W.Y. Liu, Soft Sensor Methods Study on Product Quality Based on Variable Selection, Dalian University of Technology, Dalian, 2017.

[49] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. 67 (2005) 768.

[50] Z. Zhang, Analysis of the Linear Regression Modeling for High-Dimensional Data, Tianjin University, Tianjin, 2013.

[51] L. Sun, Q. Huang, Practical Machine Learning, Posts and Telecommunications Press, Beijing, 2017.

[52] J.B. Yue, H.K. Feng, G.J. Yang, Z.H. Li, A comparison of regression techniques for estimation of above-ground winter wheat biomass using near-surface spectroscopy, Remote Sens.-Basel 10 (2018) 66.

[53] W. Yu, B.Z. Li, M.Y. Y, X.Y. Du, Building multi-objective predicting model based on artificial neural network, J. Cent. South Univ. 43 (2012) 4949–4955.

[54] L. Yang, L.Q. Hou, H.L. Li, X.Y. Xu, J.P. Liu, Regression models for energy consumption prediction in air-conditioned office building, J. Xi'an Univ. Arch. Technol. 47 (5) (2015) 707–711.

[55] B.Y. Qin, Y.Q. Pan, L.L. Yu, Z.Z. Huang, Impact of floor area on energy use intensity (EUI) of office buildings in hot-summer-and-cold-winter area of China, Build. Energy Effic. 2 (2017) 112–116.

[56] L. Zhu, Analysis of the Factors Affecting the Energy Consumption of Residential Buildings in Chengdu, Southwest Jiaotong University, Chengdu, 2017.

[57] Z. Wang, Y. Wang, R. Zeng, R.S. Srinivasan, S. Ahrentzen, Random Forest based hourly building energy prediction, Energy Build. 171 (2018) 11–25.

[58] H.W. Sun, D.W. Gui, B.W. Yan, Y. Liu, W.H. Liao, Y. Zhu, C.W. Lu, N. Zhao, Assessing the potential of random forest method for estimating solar radiation using air pollution index, Energy Convers. Manag 119 (2016) 121–129.

[59] Z.F. Cao, Study on Optimization of Random Forest, Capital University of Economics and Business, Beijing, 2014.

[60] A. Garg, K. Tai, Comparison of statistical and machine learning methods in modelling of data with multicollinearity, Int. J. Model. Identif. Control 18 (2013) 295–312.