



OPEN

Machine learning models for predicting rural residential carbon emissions and optimising spatial forms

Xu Cui¹, Yao Xu¹✉, Liang Sun²✉ & Tianning Yao³

Global warming is exacerbating its effects on ecosystems and human populations, with carbon emissions identified as the primary cause. As a crucial aspect of urban development, spatial form significantly impacts energy efficiency and carbon emissions. However, research on rural areas has been limited compared to that on metropolitan and central cities. This study investigates carbon emissions and spatial form in rural residential areas. It employs Random Forest, XGBoost, and BP Neural Network methodologies to predict carbon emissions and optimise spatial form. The findings indicate that (1) spatial form factors, including floor area ratio, number of floors, and building orientation, exhibit a strong correlation with carbon emissions; (2) the XGBoost model demonstrates superior prediction accuracy and generalization ability, achieving a reduction of over 10% in carbon emissions under the optimized spatial form; (3) optimization strategies, such as regulating the floor area ratio and minimizing the building shape coefficient, are proposed. These results provide a scientific foundation for low-carbon rural development and facilitate a green transition.

Keywords Carbon emission, Machine learning, Spatial form, Rural

In 1988, the United Nations Intergovernmental Panel on Climate Change (IPCC) expressed concerns regarding global warming^{1–3}. The Working Group I report from the Sixth Assessment, published in 2021, indicates that the global average surface temperature has risen by nearly 1 °C above pre-industrial levels and is projected to reach 1.5 °C by 2040⁴. Carbon dioxide emissions are recognised as the primary driver of global warming, leading to significant climatic changes⁵. In response to climate change and to mitigate environmental constraints, China pledged at the 21st United Nations Climate Change Conference to limit CO₂ emissions to approximately 11.7 billion tonnes by 2030 and to achieve carbon neutrality by 2060⁶.

Most current studies on carbon emissions concentrate on urban areas, emphasising building energy efficiency⁷, transport system optimisation, and energy structure transformation⁸ and have developed relatively mature prediction models and control pathways. In contrast, systematic spatial carbon emission assessment methods in rural areas are lacking due to dispersed residential land use, diverse building forms, and complex energy utilisation⁹. Most existing studies rely on regional statistical scales, making it challenging to capture the impact of spatial characteristics on carbon emissions at the micro-scale within the rural built environment. Unlike urban spatial patterns dominated by centralised energy supply and high-rise residential buildings, rural settlements are characterised by low-rise self-built dwellings, a variety of building types, a loose spatial structure, and ‘non-standardised’ and ‘diversified’ energy consumption characteristics.

Spatial form refers to the physical structure and layout of the environment, encompassing the arrangement of spaces and the diverse activities occurring within towns and cities. This concept significantly impacts the functioning of urban areas and various elements within them¹⁰. Furthermore, it is a crucial factor influencing energy efficiency and carbon emissions¹¹. Research on the correlation between spatial configuration and carbon emissions initially emerged from a macro-level perspective, concentrating predominantly on large and medium-sized urban areas. Newman et al. established a robust link between urban density, mixed-use land, and transportation energy use¹². Additionally, other scholars have identified that spatial form elements, including functional layout¹³, land use¹⁴, road system¹⁵ and floor area ratio¹⁶, significantly influence carbon emissions. Micro-level studies are typically conducted at the block and parcel level. Scholars have proposed that various indicators,

¹School of Architecture, Southwest Jiaotong University, Chengdu 611756, China. ²School of Architecture & Design, China University of Mining and Technology, Xuzhou 221116, China. ³College of Architecture & Urban Planning, Tongji University, Shanghai 200092, China. ✉email: xuyao@my.swjtu.edu.cn; sunliang@cumt.edu.cn

such as the average number of floors¹⁷ building shape coefficient¹⁸ and window-to-wall ratio¹⁹ substantially impact building energy usage by altering the internal environment of the structure. Rong P's research indicates that the energy consumption of a collection of buildings on a site cannot be determined by aggregating the energy consumption of each structure individually. It highlights that changes in the microclimate caused by mutual shading among buildings impact energy consumption. That building height and orientation within a plot are spatial form factors influencing carbon emissions²⁰. However, most existing research on spatial form and carbon emissions has focused on metropolitan areas and central cities. With the increasing policy emphasis on rural revitalisation strategies in China²¹ addressing rural issues has become increasingly crucial. To bridge this research gap in rural areas, it is essential to explore optimisation strategies for the spatial form of rural residential land use to achieve a reduction in carbon emissions.

In recent years, the development of digital and intelligent tools has made machine learning an important method for carbon emission research. It is capable of modelling nonlinear relationships, identifying key drivers and predicting emission trends based on multi-source data. Mainstream algorithms such as Random Forest²² XGBoost²³ and Neural Networks²⁴ have been widely used in carbon emission research across the fields of construction, industry and transportation due to their advantages in handling high-dimensional features and modeling complex interactions. Wen L et al. employed three machine learning models—Random Forest, the PSO algorithm, and the BP Neural Network—to forecast carbon emissions in the business sector²⁵. Wu et al. developed an innovation-driven prediction model for carbon emission intensity, introducing the XGBoost algorithm to investigate the intrinsic relationship between the pathways to carbon peaks in various countries and the variations in intellectual and physical urbanisation trajectories²⁶. Meanwhile, scholars such as Huang et al. have also focused on the nonlinear and threshold effects of building form on carbon emissions, further expanding the modelling methods that connect building spatial structure with carbon emissions²⁷.

Although numerous studies have employed machine learning methods to predict carbon emissions, most have concentrated on urban or regional scales. There is a notable scarcity of systematic modelling and optimisation analyses regarding the coupled relationship between the spatial patterns of rural dwellings and carbon emissions. Furthermore, existing studies rarely construct a microscopic, multi-dimensional feature system that encompasses spatial form, building structure, and energy use behaviour, and they often lack a refined carbon emission prediction specifically for single-family houses. This paper innovatively introduces three types of machine learning models—Random Forest, XGBoost, and BP neural network—into a multi-model cross-validation framework for rural residential carbon emissions. Additionally, it constructs a spatial form system that integrates indicators such as floor area and floor area ratio, enabling a practical model for predicting, optimising, and re-predicting the spatial form and carbon emissions in rural areas. This study not only enhances the modelling methods for rural carbon emissions but also provides scientific quantitative support for low-carbon rural planning.

Methodology

The workflow of the study

Figure 1 illustrates the procedure of this study. Spatial form indicators were developed and analysed to assess the impact of rural residential land design on carbon emissions. A machine learning model for estimating carbon emissions from rural residential properties was constructed, consisting of four phases:

Phase 1: Data collection and processing.

Relevant data on rural residential land within the study area were collected, with an emphasis on spatial form indicators such as building footprints, heights, topography, and energy consumption. The ArcGIS platform was employed to digitise the data and develop a regional spatial model. Each data point was quantified using appropriate formulas, while the dataset was cleaned and pre-processed to address missing or inconsistent values effectively.

Phase 2: Correlation analysis and regression.

Pearson correlation analysis was performed using SPSS to identify spatial form indicators significantly correlated with carbon emissions. Subsequently, single-indicator regression analyses were conducted on the selected indicators, and regression curves were plotted to visualise the relationships between each indicator and carbon emissions.

Phase 3: Machine learning prediction model construction.

Three machine learning models—Random Forest, XGBoost, and Backpropagation (BP) Neural Network—were developed on the Python platform for multiple regression analysis, utilising the selected spatial form indicators. The models were trained and validated to evaluate their predictive performance and to identify the spatial form factors with the greatest influence on carbon emissions.

Phase 4: Spatial form optimisation and carbon emission prediction.

Based on the results obtained from the machine learning models in Phase III, targeted optimisation schemes and strategies were designed. Parameters of the optimisation schemes were adjusted to simulate optimised spatial form models. Prediction outcomes before and after optimisation were compared to evaluate the effectiveness of the proposed schemes and their efficiency in reducing carbon emissions.

The study area

Macheng Township is located between latitudes 32°43' and 32°58' north and longitudes 117°08' and 117°23' east, in the central region of the Huai River in Anhui Province, China. This study focuses on Ma Township, which covers an area of approximately 5.62 km² within a total land area of 176.03 km² and has a population of 8,009 residents. According to the Chinese building thermal efficiency code, this region is classified as the “hot summer and cold winter A zone,” characterised by a transitional climate between the humid monsoon of the northern subtropical zone and the semi-humid monsoon of the southern temperate zone.

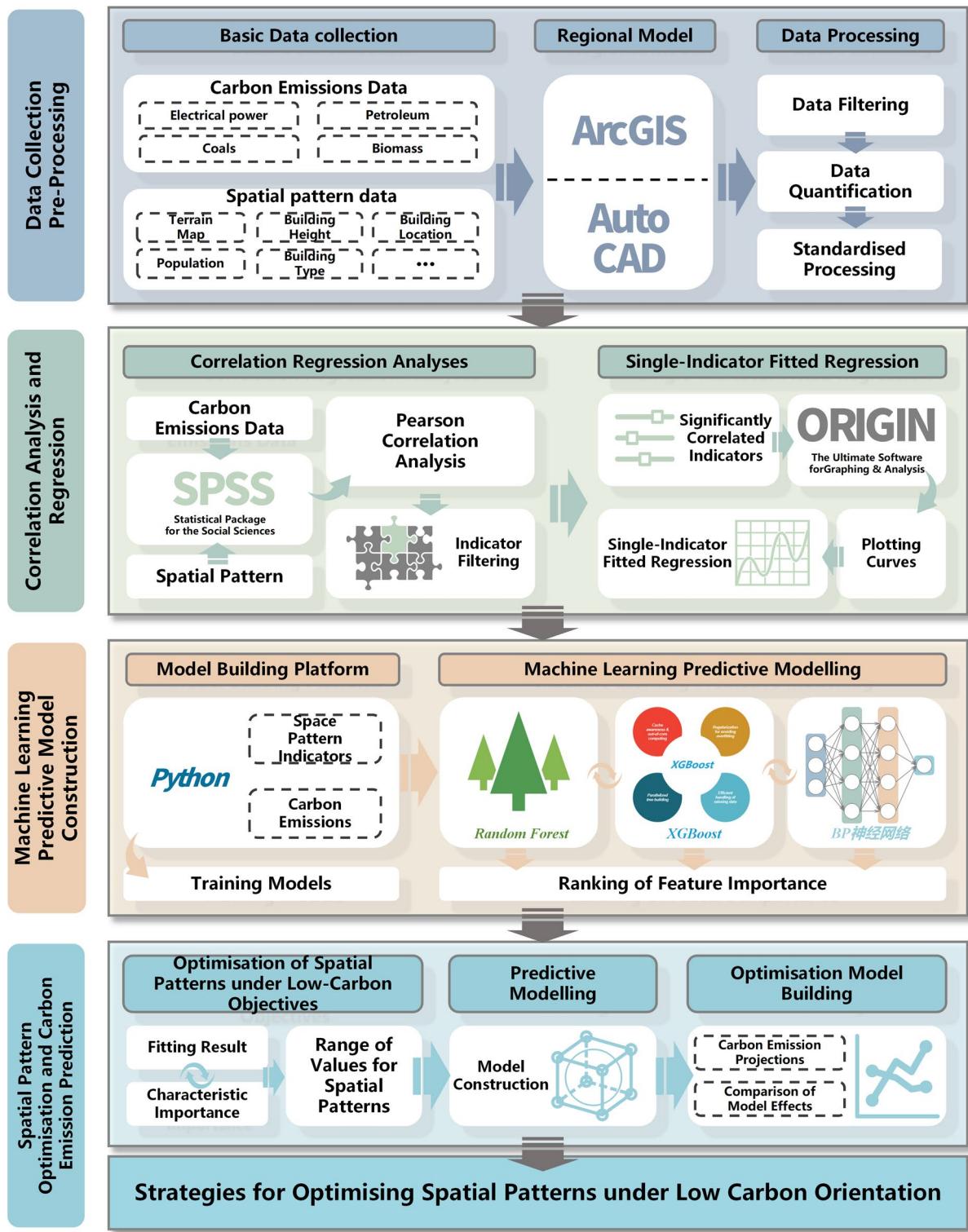


Fig. 1. The workflow of this study.

As illustrated in Fig. 2, this study gathers detailed information about the residential blocks in Macheng, including site area, building height, building area, plan dimensions, and building age. This data will support the analysis of the correlation between spatial form and carbon emissions.

The spatial unit of analysis in this study is a single household, with each household treated as an independent unit for carbon emission estimation and spatial form characterisation. This choice aims to accurately reflect the micro-distribution characteristics of carbon emissions from rural residential land, fully capturing the building



Fig. 2. Location of Macheng Township and sample distribution of rural residential areas (Created based on the standard base map downloaded from the Standard Map Service website of the Ministry of Natural Resources of China [map review number] GS(2020)4619, <https://bzdt.ch.mnr.gov.cn/>).

Data Name	Data Source
Electricity Consumption	Bengbu Power Supply Company
Natural Gas Consumption	Bengbu Xin'ao Gas Company
Liquefied Petroleum Gas Usage	Macheng Town LPG Exchange Station, Field Survey
Anthracite Coal Consumption	Field Survey
Firewood Consumption	Field Survey
2023 Administrative Boundary Map	Bengbu Natural Resources and Planning Bureau
Macheng Town Master Plan	Bengbu Natural Resources and Planning Bureau

Table 1. Summary of data sources.

scale, energy use, and spatial layout of different households. This approach provides a foundation for a detailed analysis of the factors affecting carbon emissions.

Data collection is primarily achieved through a combination of field surveys and relevant government statistics, alongside publicly available energy consumption and demographic information from local governments, ensuring the completeness and accuracy of the data. Specific data sources are listed in the following table.

Methods for the quantification of carbon emissions

Rural areas require a continuous supply of energy to function effectively, with energy consumption predominantly concentrated in three major sectors: buildings, production, and transportation²⁸. The energy utilised in buildings constitutes a significant portion of rural energy consumption, exceeding 70%²⁹. Research on carbon emissions throughout a building's life cycle can be categorised into three primary phases: construction, operation, and deconstruction³⁰. Notably, energy consumption during the operational phase is particularly critical, representing approximately 90% of the total energy consumption across the building's life cycle³¹. Consequently, this phase emerges as the primary factor in assessing construction energy use. This study evaluates the carbon emissions from rural residential land based on energy consumption during the operational phase, focusing primarily on electricity, coal, and natural gas as energy sources. Carbon emissions are quantified using a combination of direct and indirect methodologies to ensure a comprehensive assessment. Direct emissions refer to greenhouse gas (GHG) emissions resulting from the combustion of fossil fuels, including coal and natural gas. This study identifies that direct carbon emissions predominantly stem from the combustion of fossil fuels such as natural gas, liquefied petroleum gas (LPG), and anthracite coal, as well as from biomass combustion. Indirect emissions are associated with the utilisation and consumption of acquired power, heat, and steam. However, since the research area does not utilise outsourced heat or heating, only emissions derived from electricity are calculated. The computation formula is presented in Table 2.

Selection of influential spatial form indicators

Numerous studies have explored the spatial form indicators influencing carbon emissions at the site level. Baek Cheonghoon (2013) established a framework for estimating carbon emissions during the initial design phase of buildings³². Gardezi and Shafiq (2019) utilised floor area and building volume as variables to develop a predictive model for the carbon footprint during the operational phase³³. Additionally, the '3D' theory proposed by Cervero, a renowned scholar in spatial planning, has been instrumental in promoting sustainable spatial development. This theory underscores the significance of spatial form in enhancing land-use efficiency, optimising building layouts, and reducing energy consumption³⁴. It offers a comprehensive examination of the three fundamental components of spatial form: density, diversity, and design. Leng Hong (2023) proposes that low-carbon planning at the spatial level of county towns can be based on five perspectives: low-carbon layout, industry, transportation, human settlements, and ecological green space. This approach comprehensively considers the connection between land use types, land use mix, urban development density, industrial land area, accessibility of slow-moving transportation, building functions, green space ratio, and other planning indicators related to low-carbon planning³⁵. P. Depecker has also pointed out that the building shape coefficient, which reflects the ratio of the building envelope to its volume and determines the heat exchange capacity per unit volume, significantly impacts energy consumption. Furthermore, the orientation of the building, affecting solar radiation gain, ventilation, and light conditions, is a crucial factor in the design of rural dwellings³⁶.

Prior research has not specifically addressed the influence of carbon emissions from both integrated design and building perspectives. Most indicators from both macro and micro viewpoints are primarily related to energy consumption. (Table 3).

This study builds upon the '3D' theory and the contributions of various scholars to establish a foundational framework for the spatial form of rural residential land, structured around two dimensions: density elements and design elements. To minimise redundancy among indicators, the study identifies key spatial form indicators that influence carbon emissions. These indicators include the floor area ratio, degree of land mixing, building density, population density, building area, number of floors, building shape coefficient, building orientation, and construction age. Given the focus on carbon emissions from rural residential areas, the green space rate is omitted due to its negligible influence. The quantitative formulas for these specific indicators are detailed in Table 4 below.

Emission method	Formulas
Direct emissions	$E_{Fossil\ fuel} + E_{Biomass}\#(1)$
	$E_{Fossil\ fuel} = \sum_{i=1}^n (FC_i \times Car,i \times OF_i \times \frac{44}{12}) \#(2)$
	$E_{Biomass} = FBiomass \times EF_{Biomass}\#(3)$
Indirect emissions	$E_{electric} = AD_{electric} \times EF_{electric}\#(4)$

Table 2. Carbon emission formula. Note: Default values of parameters refer to the "Guidelines for the Preparation of Provincial Greenhouse Gas Inventories". Biomass only charcoal, with an emission factor of 6.0. $EF_{electric}$ is using 0.5810.

Dimension	Influential indicators in previous literature	Whether selected as an influential indicator in this study
Macroeconomics	Nature/function of the site ^{37,38,39,40}	NO
	Floor area ratio ^{16,20,41,42}	YES
	Green space ratio ^{42,43}	NO
	Degree of land mix ^{34,39,44}	YES
	Building density ^{20,39,42,44}	YES
	Population density ^{38,43,44}	YES
Microstructure	Building heights ^{42,45,46,47,48}	Replaced by the building's number of floors
	Floor area ^{37,49,50,51,52}	YES
	Building shape coefficient ^{20,48}	YES
	Building volume ^{46,48,49}	Replaced by the Building shape coefficient and Floor area ratio
	Building number of floors ^{20,37,52}	YES
	Construction age ³⁸	YES
	Building orientation ^{38,53}	YES
	Window-to-wall ratio ^{38,19,20,35,50,54}	NO

Table 3. Selection of Spatial form indicators affecting carbon emissions.

Dimension	Indicators	Formula	Descriptive
Density	Floor area ratio	$FAR = \frac{F \cdot A}{L}$	the ratio of gross floor area to lot area
	Building density	$BD = \frac{A}{L}$	the ratio of building footprint to lot area
	Degree of land mix	$DLM = \frac{-\sum P_i \ln Z_i}{\ln T_y}$	The ratio of the floor area of buildings of other functional natures to the total floor area of buildings of a single nature on a single building site.
	Population density	$PD = \frac{P_o}{L}$	population density of the area
Design	Floor area	$FA = F \cdot A_i$	represents, to some extent, the development intensity of the site
	Building number of floors	$BNF = F_i$	the natural number of floors of the building
	Building shape coefficient	$BSC = \frac{S}{V}$	a quantitative relationship between the surface area and volume of a building
	Building orientation	$BO = \frac{L_e}{R} * \cos \alpha $	the ratio of the length of the front (south) elevation of the building plan to the building's perimeter
	Construction age	$CA = C \cdot A_i$	when the building was developed and constructed

Table 4. Quantitative formulas for Spatial form indicators.

Correlation-fit analysis

Correlation analysis is a statistical technique employed to uncover the relationships between variables. This study utilises the Pearson correlation test to evaluate the extent of the linear relationship between two variables⁴⁰. By examining the magnitude of the correlation coefficient, we can identify spatial form indicators that significantly correlate with carbon emissions from construction land. Although multiple linear regression techniques are widely used in quantitative research, the incorporation of various spatial factors often increases the likelihood of encountering distinct statistical complications. Traditional multiple linear regression models face challenges in addressing the complexity and nonlinear relationships that may arise among a large number of explanatory variables⁵⁵.

Fitted regression analysis utilising a single indicator is a statistical technique employed to investigate and forecast the relationship between an independent variable (indicator) and a dependent variable. This method quantifies the effect of the independent variable on the dependent variable using a mathematical equation, thereby facilitating predictions of the dependent variable based on the value of the independent variable⁵⁶. This study developed a single-indicator fitted regression model to assess the unique influence of spatial form indicators on carbon emissions.

Machine learning prediction methods

In this study, we select three machine learning models—Random Forest (RF), XGBoost, and BP neural network—for carbon emission prediction, based on several considerations. First, RF, as an ensemble tree model, demonstrates strong resistance to overfitting and provides valuable feature importance assessments, making it well-suited for modelling nonlinear and high-dimensional features. Second, XGBoost exhibits high prediction accuracy and generalisation capability, particularly in scenarios involving small and medium-sized samples. Its iterative mechanism, based on gradient boosting, effectively captures complex relationships between variables, making it widely applicable in the fields of energy consumption and environmental modelling. Finally, the BP neural network is adept at modelling complex nonlinear systems, which is particularly useful for exploring the

potential multi-order coupling mechanisms between spatial patterns and carbon emissions. Together, these three models represent distinct algorithmic frameworks. Their combined application enhances the identification of influencing factors from multiple perspectives, improves prediction robustness, and facilitates cross-validation of results.

Random forest

Random Forest (RF) is an ensemble learning algorithm in machine learning, introduced by Breiman⁵⁷. It has emerged as one of the most widely used algorithms in this field. RF enhances the Classification and Regression Tree (CART) algorithm by utilising multiple decision trees to improve forecasting accuracy⁵⁸. Unlike traditional linear regression models, RF can effectively identify and represent complex nonlinear interactions among predictor variables, making it exceptionally versatile and powerful for a variety of applications⁵⁹.

Random Forest is an integrated learning model constructed from numerous decision trees. By combining the predictive outputs of each tree, the model achieves more accurate and stable predictions. Its notable advantages include a significant reduction in the risk of overfitting when handling complex datasets with many features, as well as improved robustness in managing missing data.

For any sample X with P sub-models, P predicted values will be generated. Assuming that the expected value of the kth submodel is \hat{y}_K , the overall prediction \hat{y}_E will be obtained by simple averaging⁶⁰:

$$\hat{y}_E = \frac{1}{P} \sum_{k=1}^P \hat{y}_K \quad (5)$$

XGBoost

XGBoost, which stands for eXtreme Gradient Boosting, is an ensemble learning technique that utilises decision trees. It integrates a powerful linear model solver with tree-based learning methodologies, accommodating various objective functions such as regression, classification, and ranking. XGBoost employs sophisticated optimisation methods, including modifications to sample weights, regularization, and management of model complexity, to enhance forecast accuracy. Its impact is widely recognised, particularly in machine learning and data mining tasks, where it consistently yields exceptional results⁶¹.

For a dataset of n entries across m dimensions, the XGBoost model can be formulated as follows:

$$y_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in F \quad (i = 1, 2, \dots, n) \quad (6)$$

Among them, $F = [f(x) = w_{q(x)}] \quad (q : R^m \rightarrow (1, 2, \dots, T), w \in R^T)$ represents the CART decision tree structures. Constructing the XGBoost model involves optimising its parameters to minimise the objective function, thereby obtaining the best possible model. The objective function consists of two parts: the loss function L and the regularization term Ω , expressed as:

$$Obj = L + \Omega \quad (7)$$

$$L = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (8)$$

$$\Omega = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (9)$$

BP neural network

In 1986, Rumelhart, Geoffrey Hinton, and their colleagues introduced the BP neural network, which has attracted significant attention across various domains, including pattern recognition, function approximation, and data mining. The BP neural network effectively learns and simulates complex nonlinear mapping relationships by fine-tuning the weights among neurons across different layers and utilising the backpropagation error mechanism.

During the forward propagation phase, input samples are first passed through the input layer to the hidden layer. Subsequently, these sample data are processed in the hidden layer using a nonlinear activation function $\varphi(x)$ for transformation before being further passed to the output layer. The formula for input net_i for the implicit layer is as follows:

$$net_i = \sum_{j=1}^m w_{ij} x_j + \beta_i \quad (10)$$

The output of the implicit layer y_i :

$$y_i = \varphi(net_i) = \varphi\left(\sum_{j=1}^m w_{ij} x_j + \beta_i\right) \quad (11)$$

Input net_k for the output layer:

$$net_k = \sum_{i=1}^q w_{ki} y_i + b_k = \sum_{i=1}^q w_{ki} \varphi\left(\sum_{j=1}^m w_{ij} x_j + \beta_i\right) + b_k \quad (12)$$

The output of the output layer G_k :

$$G_k = \lambda(net_k) = \lambda\left(\sum_{i=1}^q w_{ki} y_i + b_k\right) = \lambda\left(\sum_{i=1}^q w_{ki} \varphi\left(\sum_{j=1}^m w_{ij} x_j + \beta_i\right) + b_k\right) \quad (13)$$

Model evaluation

R^2 (coefficient of determination), MSE, and RMSE indicators were used to measure the three models' accuracy comprehensively.

- R^2 (coefficient of determination).

R^2 referred to as the goodness-of-fit metric, ranges from 0 to 1. A value of R^2 approaching 1 indicates that the model fits the target variable more effectively, accounting for a greater share of the variance in that variable.

$$R^2 = 1 - \frac{\sum_i (y_i - \bar{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (14)$$

- MAE (Mean Absolute Error).

MAE quantifies the mean magnitude of absolute deviation between the predicted outcomes of a model and the actual values. A smaller MAE value indicates a lesser average prediction error, thereby reflecting a higher level of prediction accuracy. The formula for calculating MAE is as follows:

$$MSE = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y}_i)^2 \quad (15)$$

- RMSE (Root Mean Square Error).

RMSE is defined as the square root of the mean squared deviation between expected and actual values. It quantifies the divergence of projected values from actual outcomes and is particularly sensitive to outliers. The equation for MSE is as follows:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \bar{y}_i)^2} \quad (16)$$

Results

Distribution of carbon emissions

This study encompasses a sample of approximately 1,800 households. Due to the inability of government departments to provide data on biomass combustion, this information was exclusively obtained through research. Among the sampled households, 340 utilise biomass combustion, while the remainder rely on electricity. The sample was drawn using a stratified random sampling method, which captures various topographic features and types of housing sites, including centralised settlements, natural villages, and self-built homes by farmers. Data sources comprised field questionnaires, government energy use statistics, and remote sensing data overlays. For data processing, techniques such as imputing missing values, eliminating outliers, and cross-referencing data with local energy management departments were employed to ensure data quality.

Based on field research and data from government departments, the primary energy sources for rural residential areas in Macheng Town include natural gas, liquefied petroleum gas, electricity, coal, and charcoal. Carbon emissions were calculated using the formula specified in subsection 2.3. The carbon emission statistics for rural residential areas in Macheng Town for the year 2022 are summarised in Table 5. The total carbon emissions reached 3775.34 tCO₂, with electricity consumption constituting the largest share at 88.87% of the total emissions.

Correlation analysis

The results presented in Fig. 3 allow for several conclusions to be drawn. The Pearson correlation coefficients among the independent variable indicators are all below 0.8, indicating a lack of high correlation among them. This finding suggests that multicollinearity is not an issue, allowing for the retention of all selected indicators without exclusion. Among the spatial form indicators, a significant correlation exists between eight indicators and the carbon emissions of rural residential land. Specifically, the indicators of building area, number of floors, building orientation, and shape coefficient are significantly correlated at the 0.01 level. Additionally, the indicators of floor area ratio, building density, population density, and construction age are significantly correlated at the 0.05 level.

To investigate potential nonlinear relationships between spatial form indicators and carbon emissions, this study performed curve fitting analyses. Metrics such as the probability value (Prob > F) to assess the significance of the fit and the coefficient of determination (R^2) to evaluate the explanatory power of the trend line were compared. The optimal results are presented in Fig. 4; Table 6.

Direct emission/tCO ₂				Indirect emission/tCO ₂	
petroleum	liquefied petroleum gas	anthracite	charcoal	electrical power	Total/tCO ₂
143.33	52.49	223.16	1.08	3355.28	3775.34

Table 5. Statistics on carbon emissions.

The correlation analysis between spatial form indicators and carbon emissions

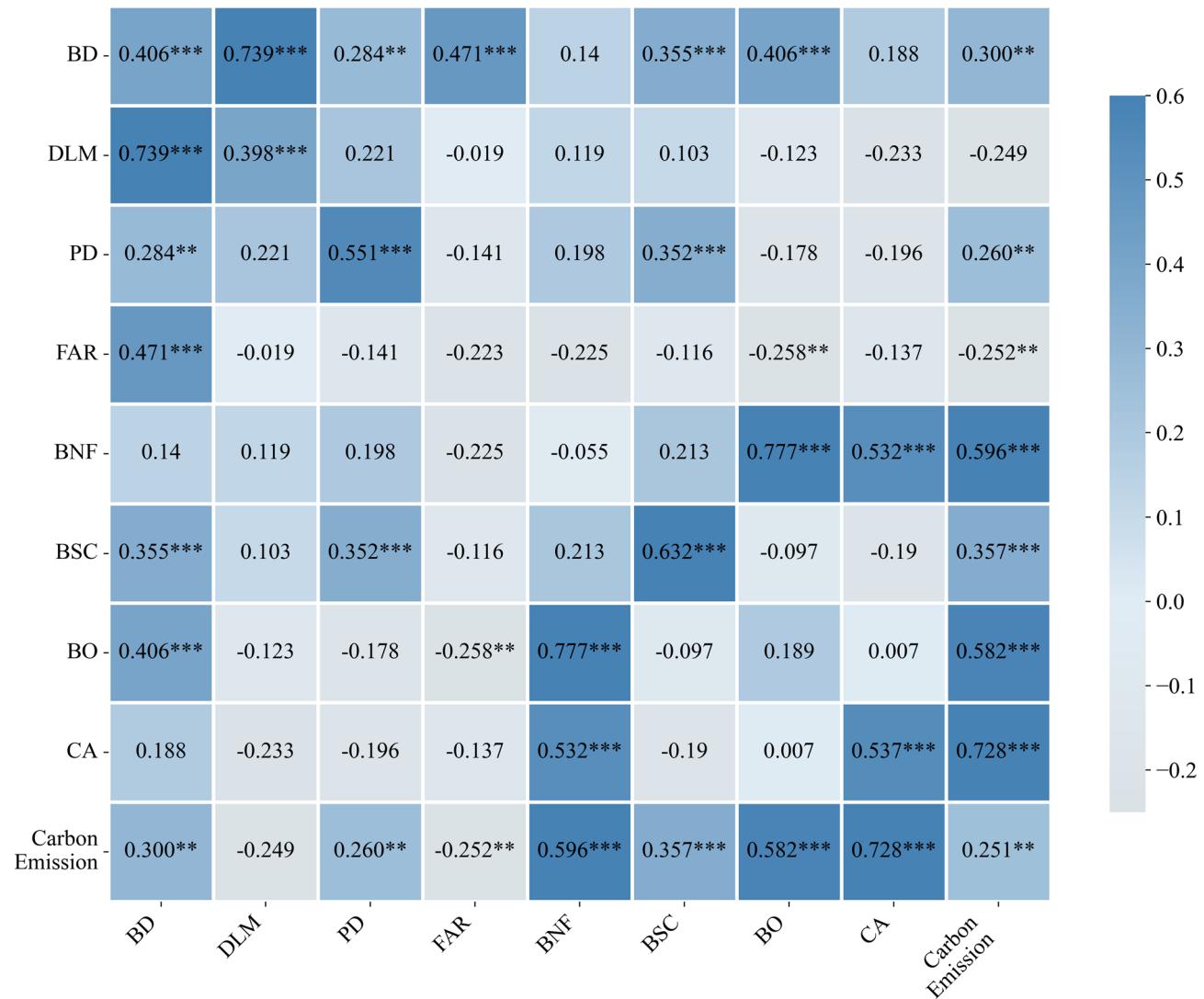


Fig. 3. Correlation analysis between spatial form indicators and carbon emissions.

Machine learning

The multivariate regression relationship between spatial form and carbon emissions was further analysed using machine learning regression models. Figure 5 illustrates the regression graphs comparing the predictive performance of Random Forest, XGBoost, and Backpropagation neural networks. The horizontal axis represents actual carbon emissions, while the vertical axis displays the expected values produced by the models. Data points close to the red diagonal line indicate instances where the projected values align with the actual values, demonstrating the accuracy of the models. The chart shows that the projected values from all three models closely correspond with the actual values, indicating a superior overall fitting ability.

As shown in Table 7, the model demonstrates a superior fit to the data when the MSE, RMSE, and MAE values are minimised. Furthermore, an R² Score approaching 1 signifies a better model fit. The accuracy metrics indicate that the Random Forest, XGBoost, and BP neural network models all achieve R² Scores exceeding 0.98 in predicting carbon emissions. Additionally, both the MAE and RMSE values for the test and training sets remain below 0.1, illustrating that all three models exhibit exceptional fitting performance.

Feature importance is a metric employed in machine learning to quantify the influence of each feature on model predictions. This metric can be derived from an analysis of the model's structure and the decision-making process. To visualise the features that play a critical role in model predictions, the spatial representation of the three models is ranked according to their importance, as illustrated in Fig. 6.

From the figure, it can be observed that the top six spatial form indicators affecting carbon emissions, as derived from the random forest model, are FA, FAR, BSC, PD, BD, and BO, listed in descending order. The important indicators identified by XGBoost, also in descending order, are FA, FAR, PD, BD, CA, and BNF. Finally, the indicators derived from the BP neural network are FA, BSC, BNF, BD, CA, and BO.

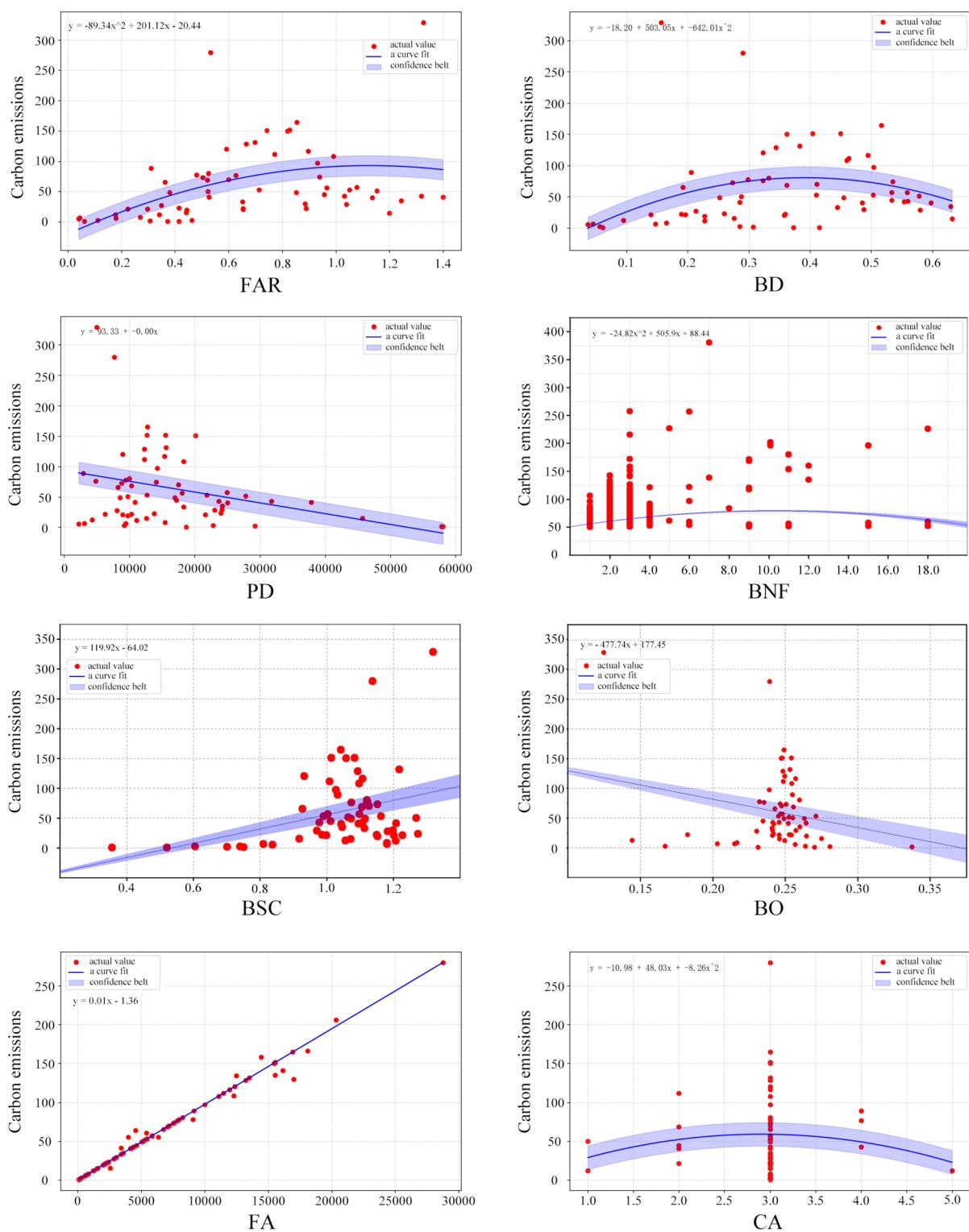
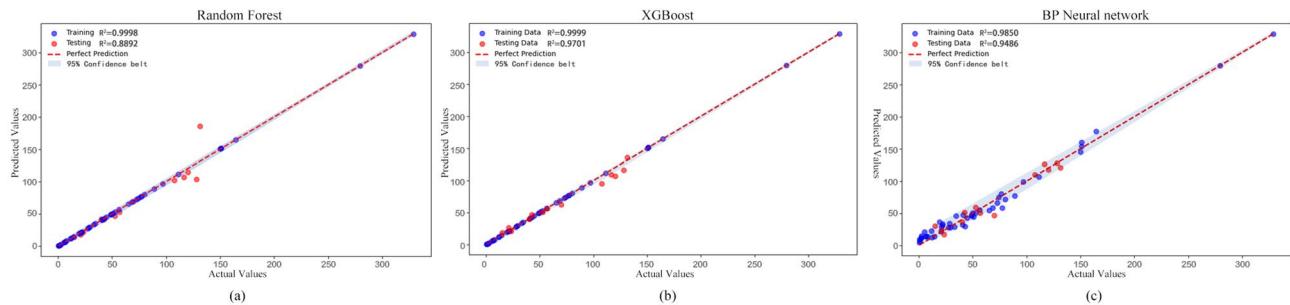


Fig. 4. Regression curve of spatial form indicators fitted to carbon emissions.

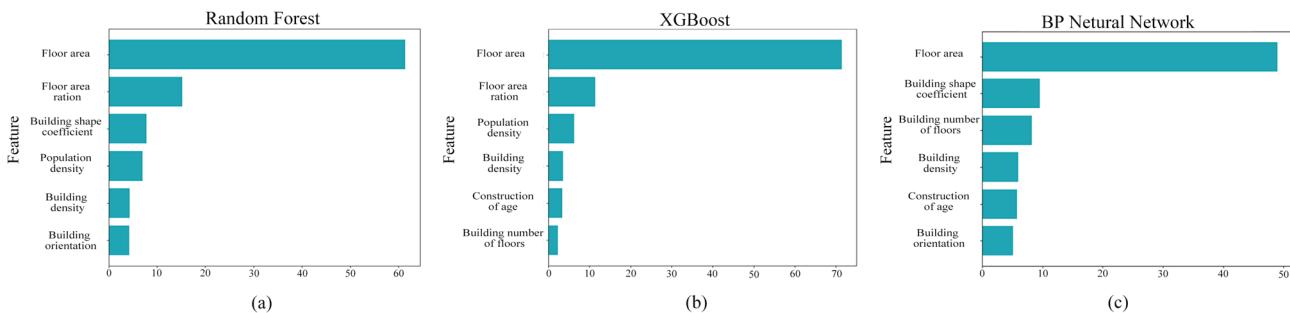
Optimisation strategies

We are selecting the reduced interval for the spatial form indicators and the optimal fitting results for carbon emissions. These are connected with the values of the control indicators required by various norms. This allows us to obtain the suggested range of values for the spatial form indicators under a low-carbon orientation, as shown in Table 8 below.

Dimension	Indicator	Best-fit curve	Extreme/trend	R ²	Prob> F
Density	FAR	Inverted "U" Curve	Maximum value: 1.10	0.5752	0.031
	BD	Inverted "U" Curve	Maximum value: 0.39	0.6929	0.004
	PD	Linear negative correlation	Decreasing	0.6228	0.002
Design	BNF	Inverted "U" Curve	Maximum value: 10.19	0.3367	0.045
	BSC	Linear Positive Correlation	Increasing	0.6153	0.003
	BO	Linear negative correlation	Decreasing	0.5518	0.006
	FA	Linear Positive Correlation	Increasing	0.9857	0.006
	CA	Inverted "U" Curve	Maximum value: 2.95	0.5458	0.008

Table 6. Results of optimal fitting of Spatial form indicators to carbon emissions.**Fig. 5.** Regression plot of machine learning prediction models (a) Random Forest model; (b) XGBoost model; (c) BP neural network model.

Model	Training R ²	Test R ²	Training MAE	Test MAE	Training RMSE	Test RMSE
RF	0.9998	0.8892	0.0006	0.0102	0.0001	0.0998
XGBoost	0.9999	0.9701	0.0004	0.0589	0.0009	0.0725
BP	0.9850	0.9486	0.0652	0.0748	0.0081	0.0951

Table 7. Machine learning prediction model accuracy and fitting effects.**Fig. 6.** Importance of machine learning model features (a) Random Forest model; (b) XGBoost model; (c) BP neural network model.

Verification

Six representative sites within the parcel were selected to optimise and validate the carbon emission prediction model, ensuring data diversity and completeness. During the optimisation process, priority was assigned to the spatial form indicators identified as most significant through the machine learning analysis. These indicators were adjusted moderately within the suggested value ranges derived from fitting analysis and relevant specifications.

Based on the key influencing factors identified by the model, this study proposes several spatial form optimisation paths with a policy basis. Specific strategies include: ① Reducing the number of fragmented buildings and integrating land use to decrease building density and shape coefficients; ② Moderately increasing the number of building floors to improve the floor area ratio within the scope of the code; ③ Guiding the development

Spatial form indicators	Typology	Regulatory requirement	Range of values for low-carbon recommendations
FAR ^{①②}	new building site	1.0-2.2	common residential 1.1–1.5 resettlement apartment 1.5–1.8
	rural homestead	0.8–1.2	0.8–1.1
BD ^{①②}	new building site	common residential ≤ 35%	common residential ≤ 35%
		resettlement apartment ≤ 32%	resettlement apartment ≤ 32%
	rural homestead	≤ 40%	≤ 39%
PD ^③	land area 130–200hm ²	38,400 – 50,000 people/km ²	38,400 – 50,000 people/km ²
	land area 32–50hm ²	30,000–50,000 people/km ²	30,000–50,000 people/km ²
	land area 8–18hm ²	27,800 – 66,700 people/km ²	27,800 – 66,700 people/km ²
	land area 2–4hm ²	25,000–75,000 people/km ²	25,000–75,000 people/km ²
BNF ^④	new building site	6–18 floors	6–18 floors
	rural homestead	1–6 floors	1–6 floors
BSC ^⑤	building number of floors ≤ 3 floors	≤ 0.60	≤ 0.60
	building number of floors > 3 floors	≤ 0.40	≤ 0.40
BO ^⑥	-	The south orientation of the building shall be 15° southwest to 15° southeast.	Predominantly south-facing buildings
FA ^{①②}	low rise building(1–3 floors)	> 10,000 m ²	> 10,000 m ²
	multi-story building(4–9 floors)	> 80,000 m ²	> 80,000 m ²
	high-rise building(10–18 floors)	> 300,000 m ²	> 300,000 m ²
CA ^⑦	-	-	Low-carbon transformation of old residential areas

Table 8. Suggested range of values for Spatial form indicators under low-carbon orientation. Data Source:

① General Rules for Detailed Control Planning of Bengbu City (2020 Edition) ② Small Town Planning and Design ③ Planning and Design Standards for Urban Residential Areas GB50180-2018 ④ Opinions on Strengthening Green and Low-Carbon Construction of County Cities ⑤ General Specification for Building Energy Efficiency and Renewable Energy Utilisation GB55015-2021.

of building forms towards regularity to minimize the area of external enclosure structures and reduce energy consumption; and ④ Increasing the population density in new projects to promote intensive management of the energy system. The optimisation proposal is based on the ‘Urban Residential Area Planning and Design Standards’, ‘General Rules for Detailed Control Planning of Bengbu City (2020 Edition)’, and other norms to ensure that the spatial intervention design is conducted under the premise of legal compliance.

The trained regression prediction models derived from three types of machine learning were employed to estimate carbon emissions based on the optimised spatial form indicators. These predictions were subsequently compared with the current state to calculate the percentage change in carbon emissions. The details of the optimised three-dimensional spatial models and their corresponding carbon emissions are illustrated in Fig. 7.

The above Figure shows that rural residential land exhibits significant carbon reduction effects. The average carbon emissions, as predicted by the RF, XGBoost, and BP neural network models, are reduced by 10.94%, 14.49%, and 12.12%, respectively.

Discussion

Discussion of research findings

This study investigates the impact of spatial form on carbon emissions from rural residential land utilising machine learning techniques. The results of single-indicator fitting regression indicate that:

Density Dimension: Lower floor area ratios and building densities are associated with reduced carbon emissions, which is consistent with previous research⁶². However, excessively reducing these indicators is impractical. Instead, maintaining the floor area ratio within a specific range while moderately increasing building density can effectively contribute to lowering carbon emissions¹⁷. Furthermore, enhancing population density can facilitate emission reductions, as a higher living density encourages the sharing of spatial resources, thereby improving energy efficiency and decreasing emissions, as highlighted in the study by D. Timmons⁶³.

To further compare the performance of the three models, we evaluated their prediction accuracies on both the training and test sets. The results indicate that all three models exhibit strong fitting capabilities, with the XGBoost model achieving the highest R² value on the test set. This finding suggests that XGBoost demonstrates superior generalisation ability when handling small to medium-sized sample data. In terms of variable importance interpretation, the three models rank certain indicators slightly differently, primarily due to the distinct mechanisms of their algorithms. Random Forest and XGBoost, as tree-based models, assess variable importance based on the gain generated at the split nodes, which tends to favour variables with clear demarcation points in the local region. In contrast, the BP neural network relies on the back-propagation of global error. It comprehensively measures the weights of each variable throughout the overall fitting process, thereby better reflecting the synergistic effects among variables. Consequently, the differences in the ranking of certain variables across the models do not imply that their influences diverge; rather, they highlight the

NO	Photos of the current situation	Current 3D model	Optimized 3D model	Current Carbon Emissions/tCO ₂	Optimizing Carbon Emissions/tCO ₂			Magnitude of increase/decrease/%		
					RF	XG	BP	RF	XG	BP
1				28.75	23.33	26.77	28.33	Reduce 18.85	Reduce 6.89	Reduce 1.46
2				29.49	24.87	23.20	22.76	Reduce 15.68	Reduce 21.33	Reduce 22.82
3				73.92	66.31	58.04	62.03	Reduce 10.29	Reduce 21.48	Reduce 16.08
4				116.35	102.54	95.15	90.22	Reduce 11.87	Reduce 18.23	Reduce 22.45
5				279.68	268.43	254.78	271.28	Reduce 4.02	Reduce 8.91	Reduce 3.00
6				328.69	312.54	295.37	308.67	Reduce 4.91	Reduce 10.14	Reduce 6.91

Fig. 7. Comparison of carbon emissions before and after spatial form optimisation.

models' varying capacities to recognise the relationships among variables from different perspectives. Design Dimension: Consistent with previous findings, an increase in the average number of building floors reduces the building shape coefficient, thereby decreasing heat loss through the building envelope⁶⁴. Buildings with south-facing elevations receive the most solar radiation⁶⁵ and smaller floor areas minimise the surface area that absorbs solar radiation⁶⁶. Incorporating energy-efficient designs and HVAC systems in new buildings further reduces energy consumption for cooling and heating⁶⁷.

The correlation between carbon emissions and the number of building floors exhibits a pattern of initial increase followed by a decline, diverging from findings in other studies. This variance emphasises the intricate relationships between spatial form and carbon emissions.

This study represents the inaugural effort to thoroughly implement machine learning methodologies in optimising spatial form and predicting carbon emissions in rural residential land, successfully identifying a machine learning prediction model suitable for the study area and addressing a gap in current rural research.

Research limitations and future directions

This study has achieved preliminary results in the optimisation of spatial patterns and the prediction of carbon emissions for rural residential land; however, it faces limitations in the following aspects:

(1) Insufficient regional representation and limited data sources. The data in this study primarily originate from typical townships in Anhui Province, which provides some regional representativeness. However, due to the diversity of rural areas in China, the universality and applicability of the model still require further testing. Future research should broaden the scope of data collection to encompass rural areas with varying geographic locations, development stages, and spatial patterns to enhance the model's adaptability and generalizability.

(2) The dimensions of the variables are relatively narrow, as social, environmental, and behavioural factors have not been systematically incorporated. The current study focuses on the impact of spatial pattern indicators on carbon emissions, but has not yet fully considered multi-dimensional driving factors such as residents' behaviour, socio-economic activities, and climate variations. Future work could incorporate multi-source data, including household income, energy preferences, and temperature changes, to develop a more comprehensive model explaining the impact mechanisms of carbon emissions. (3) Lack of empirical cases to support the optimisation effect. Although this study demonstrates that spatial optimisation can yield a carbon emission reduction potential of 10%–14% through model predictions, it lacks verification against actual low-carbon rural construction cases. This absence hinders a comprehensive assessment of the real policy value of the strategy. In future research, we will actively engage with local pilot projects in low-carbon villages and towns, promote the integration of optimisation strategies with field planning projects, and explore a closed-loop approach of 'simulation-verification-implementation' to enhance the practical applicability of the model results.

(4) The implementation of optimisation strategies encounters several practical challenges. In reality, factors such as construction costs, land use policies, and residents' living preferences may constrain improvements in building energy efficiency and spatial layout adjustments. Future research should incorporate financial incentive data, policy and regulatory frameworks, and resident acceptance surveys to evaluate the feasibility and adaptability of optimisation measures comprehensively. Moreover, promoting the alignment between strategic interventions and community needs through participatory planning is essential.

(5) The application of technology can be further deepened and expanded. Future efforts should focus on the enhanced integration of remote sensing and GIS technologies to improve dynamic monitoring of spatial patterns. Concurrently, advanced algorithms, including deep learning and ensemble learning, should be explored to increase the accuracy of predictive models. With the rising adoption of rural renewable energy sources such as solar power and biomass gasification, subsequent studies are expected to gradually integrate these factors into the carbon accounting system to improve the model's foresight and comprehensiveness.

Conclusions and recommendations

Conclusions of the study

Based on the fitting analysis, model predictions, and optimisation strategies, ①several key conclusions are drawn regarding the relationship between spatial form and carbon emissions from rural residential land: BA, BNF, BO, BSC, FAR, BD, PD, and CA. These indicators significantly influence carbon emissions and are critical for understanding and optimising spatial form for carbon reduction. ②The impact mechanisms include: a moderate increase in FAR and BD enhances land use efficiency and facilitates centralised energy management; however, excessive density may lead to issues with ventilation and lighting, ultimately increasing energy consumption. The relationship between BNF and energy consumption follows an inverted 'U' shape: low-rise buildings with a significant envelope ratio are more efficient, while high-rise buildings in rural areas often suffer from technological shortcomings, resulting in increased energy consumption. Additionally, a lower BSC correlates with reduced heat loss per unit volume, highlighting the energy-saving advantages of compact building structures. Conversely, buildings with large BA but non-compact shapes experience a significant increase in carbon emission intensity.

In the carbon emission forecast model based on machine learning: ③Regarding model prediction accuracy, the fitting performance of the three models—random forest, XGBoost, and BP neural network—is excellent; ④Concerning the importance of features, FAR and BD in the density dimension are identified as essential indicators influencing carbon emissions, while FA and BSC in the design dimension are also critical indicators; ⑤In the optimized prediction model, the spatial form indicators predicted by the machine learning models demonstrate an average reduction in carbon emissions of over 10% across all three models; ⑥The comparative analysis indicates that the XGBoost model demonstrates enhanced efficacy in minimizing carbon emissions in the test set. With the optimized prediction model, following the prediction of improved spatial form indicators, all three models exhibit improved carbon reduction impacts, with average emissions decreased by over 10%; ⑦Furthermore, the comparison indicates that the XGBoost model achieves the highest R^2 score and exhibits the best performance in error metrics on the test set, suggesting its strong generalization capability to unknown data. Additionally, after optimising the spatial form in carbon emission predictions, the XGBoost model predicts the most significant average reduction in carbon emissions. Consequently, XGBoost is more suitable for forecasting carbon emissions in rural residential regions.

Finally, combining the "dual-carbon" goal with the background of the rural revitalisation strategy, this study proposes quantitative spatial optimisation strategies aimed at reducing carbon emissions. These strategies offer quantifiable technical paths and planning support for "intensive use and green transformation" in rural construction, demonstrating strong policy relevance and potential for application and promotion. This study also provides quantifiable technical paths and planning support for "intensive utilisation and green transformation" in current rural construction, highlighting its strong policy suitability and potential for application and promotion.

Optimisation strategies

Combining the results of the model with the characteristics of the built environment in rural areas, the following operational spatial optimisation strategies are proposed to support China's goal of achieving 'carbon peaking and carbon neutrality' and to facilitate the green and low-carbon transformation of rural construction.

Control of floor area ratio (FAR) and building density (BD): Given that FAR and BD exhibit an inverted U-shaped relationship with carbon emissions, it is recommended that the control values be maintained within a reasonable median range. This approach aims to prevent the increase in carbon emissions that can result from either excessive density or insufficient density. Therefore, it is recommended that the floor area ratio for newly constructed ordinary residential buildings be maintained between 1.1 and 1.5, with a building density of 35%. For newly built resettlement houses, the floor area ratio should be set between 1.5 and 1.8, accompanied by a building density of 32%. Additionally, the floor area ratio for rural residential land should be controlled within the range of 0.8 to 1.1, with a maximum building density of 39%.

This study focuses on optimising body shape design and building height. It proposes a moderate increase in the number of building floors while adhering to rural construction height regulations. Specifically, residential buildings should not exceed six floors, and resettlement residences should be limited to fewer than eighteen floors. Additionally, strategies to reduce the body shape coefficient will be explored. This can be achieved by implementing a design strategy that involves multiple units in series to reduce the area of external walls, expanding the building's depth, and designing regular building shapes to simplify the perimeter⁶⁸ to make the building more concentrated and compact.

These strategies aim to create a more concentrated and compact building structure. Furthermore, promoting low-carbon retrofitting of older buildings is crucial, as studies have shown that the carbon intensity of homes built before 2000 is significantly high. It is recommended that priority be given to energy-efficient upgrades for this type of building, including: It is recommended that older settlements undergo decarbonization through the enhancement of the building envelope⁶⁹ the optimization of the heating system⁷⁰ and the improvement of the lighting configuration⁷¹.

To enhance the feasibility of strategy implementation, it is recommended to leverage policy support initiatives such as 'beautiful villages' and 'livable farmhouses' to promote energy-saving renovations and financial subsidies. Additionally, improving land use efficiency can be achieved through centralised layouts and shared facility designs. Introducing participatory planning involving residents will foster community recognition and adaptation to the spatial optimisation strategies, ensuring their practicality. Emphasising intensive land use through centralised layouts and shared facilities, alongside participatory planning, will further enhance the community's acceptance and adaptability, ultimately ensuring that the strategies are both practicable and feasible.

<i>EDirect emissions</i>	direct emissions
$E_{Fossil\ fuel}$	the emission of fossil fuel combustion
FC_i	the consumption of fossil fuel
$C_{ar,i}$	the carbon concentration of the acquired base ingredient of fossil fuel
OF_i	the carbon oxidation rate of fossil fuel
44/12	the ratio of the relative molecular mass of carbon dioxide to that of carbon
i	the code name of the fossil fuel type
$E_{Biomass}$	biomass combustion emissions
$F_{Biomass}$	biomass fuel consumption
$EF_{Biomass}$	biomass fuel combustion emission facto
<i>EIndirect emssions</i>	indirect emissions
$E_{electric}$	the emissions from purchased electricity use
$AD_{electric}$	the purchased electricity use
$EF_{electric}$	the grid emission factor
FAR	Floor area ratio
BD	Building density
DLM	Degree of land mix
PD	Population density
FA	Floor area
BNF	Building number of floors
BSC	Building shape coefficient
BO	Building orientation
CA	Construction age
Fl	the number of floors
L	the lot area
A	the building footprint
Ty	the number of building types
Z_i	the percentage of floor area of the ith type
Po	the population
i	the ith building
S	the building exterior area
V	the building volume
Le	the length of the building's south-facing façade
R	the building's perimeter
α	the building's south-facing angle to the horizontal
X	sample size
P	number of submodels
y_E	final predictions of the integrated model
k	from 1 to The index of P
y_E	the predicted value of the kth submodel for the sample
F	the set of CART decision tree structures
q	the tree structure of samples mapped to leaf nodes
T	the number of leaf nodes, and w is the real fraction of leaf nodes
Ω	the model complexity function term
γT	the L1 regular term
$\frac{1}{2} \lambda \sum_{j=1}^T w_j^2$	the L2regualr term
j	the input layer contains j nodes
x	vector
w_{ij}	the weight from the jth input node to the ith hidden layer node
βi	the threshold of the ith hidden layer node
$\varphi(x)$	the activation function of the layer
k	the output layer have k nodes
w_{ki}	the weight from the ith implicit layer node to the kth output node
b_k	the threshold value of the kth output node
$\lambda(x)$	the activation function of the output layer
G_k	stands for the output result of the kth output node
Continued	

<i>EDirect emissions</i>	direct emissions
R^2	coefficient of determination
MAE	mean absolute error
$RMSE$	root mean square error
$y_i - \bar{y}_i$	the true-predicted value of the test set

Table 9. Variable Annotation Table.**Data availability**

The datasets used during the current study are available from the corresponding author on reasonable request.

Appendix

Table 9.

Received: 7 January 2025; Accepted: 26 August 2025

Published online: 26 September 2025

References

- Change, C. the IPCC scientific assessment. Contribution of Working Group I to the First Assessment Report of the Intergovernmental Panel on Climate Change, 365. (1990).
- Council, N. R. et al. *Abrupt Impacts of Climate Change* (National Academies, 2013).
- Houghton, J. T. et al. *Climate Change 2001: the Scientific Basis* Vol. 881 (Cambridge university press Cambridge, 2001).
- Lynn, J. & Peeva, N. Communications in the ipcc's sixth assessment report cycle. *Clim. Change.* **169** (1), 18 (2021).
- Shakun, J. D. et al. Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation. *Nature* **484** (7392), 49–54 (2012).
- Contributions, I. N. D. *China-intended Nationally Determined Contribution* (INDC, 2015).
- Seyedabadi, M., Abolhassani, S. & Eicker, U. Developing a Systematic Framework for Integrating Life Cycle Carbon Emission Assessment in Urban Building Energy Modeling. *Building and Environment*. (2024). <https://doi.org/10.1016/j.buildenv.2024.111662>
- Wei, C. et al. Analysis of carbon emissions in urban Building sector using multi-influence model. *J. Clean. Prod.* <https://doi.org/10.1016/j.jclepro.2023.139130> (2023).
- Wu, Y., Wang, Z., Zhu, X., Yu, H. & Jia, S. Construction and Spatial-Temporal characteristic analysis of the carbon atlas of Low-Carbon villages in the Yangtze river delta. *J. South. Archit.* <https://doi.org/10.33142/jsa.v1i1.10431> (2024).
- Kahn, M. E. *Green Cities: Urban Growth and the Environment* (Rowman & Littlefield, 2007).
- Yang, L., Li, G., Lin, Y. & Ye, L. Progress and Prospect on Relationship Research between Urban Form and Carbon Emission. *Urban Development Studies*, 18(02), 12–17 + 81. (2011). Retrieved from https://kns.cnki.net/kcms2/article/abstract?v=VKFFl0Cm57ZUlXkJYf8kolj_6vdKWsLfbmJEMDVHcjw3Ps4Op5dPUT_VS55zvU6GXWlv74EXCVNIXu9KIGsI0Bboh7ARC--xsnMjXVWoZg7C0dbFivAiX9y2jioz0ZrNMdOwwM5oyXnMLMGpgfQ==&uniplatform=NZKPT&language=CHS
- Newman, P. & Kenworthy, J. Sustainability and cities: overcoming automobile dependence. (1998).
- Wang, X. et al. Carbon emission accounting and Spatial distribution of industrial entities in Beijing—Combining nighttime light data and urban functional areas. *Ecol. Inf.* **70**, 101759 (2022).
- Chen, S., Haase, D., Qureshi, S. & Firoozaei, M. K. Integrated land use and urban function impacts on land surface temperature: implications on urban heat mitigation in Berlin with eight-type spaces. *Sustainable Cities Soc.* **83**, 103944 (2022).
- Zhang, L., Chen, H., Li, S. & Liu, Y. How road network transformation May be associated with reduced carbon emissions: an exploratory analysis of 19 major Chinese cities. *Sustainable Cities Soc.* **95**, 104575 (2023).
- Javanroodi, K., Mahdavinejad, M. & Nik, V. M. Impacts of urban morphology on reducing cooling load and increasing ventilation potential in hot-arid climate. *Appl. Energy* **231**, 714–746 (2018).
- Liu, K. et al. Impact of urban form on Building energy consumption and solar energy potential: A case study of residential blocks in jianhu, China. *Energy Build.* **280**, 112727 (2023).
- Carpio, M. & Carrasco, D. Impact of shape factor on energy demand, co2 emissions and energy cost of residential buildings in cold oceanic climates: case study of South Chile. *Sustainability* **13** (17), 9491 (2021).
- Shaik, S., Gorantla, K., Ghosh, A., Arumugam, C. & Maduru, V. R. Energy savings and carbon emission mitigation prospective of building's glazing variety, window-to-wall ratio and wall thickness. *Energies* **14** (23), 8020 (2021).
- Rode, P., Keim, C., Robazza, G., Viejo, P. & Schofield, J. Cities and energy: urban morphology and residential heat-energy demand. *Environ. Plan.* **41** (1), 138–162 (2014).
- Xi, J. Hold High the Great Banner of Socialism with Chinese Characteristics and Strive in Unity to Build a Modern Socialist Country in All RespectsReport to the 20th National Congress of the Communist Party of China. Gazette of the State Council of the People's Republic of China(30), 4–27. (2022). Retrieved from https://kns.cnki.net/kcms2/article/abstract?v=VKFFl0Cm57aqP_W55WGrgfsmC25KH_c64Vx_Yojs5YxiEzeV_GLUcehsNNkUEDgiEQW8MAgzbPwYUyKCcEPtFLDm_3Uzs5KDmMvCmvSXQhCt0qSoOxej5ML_aAf4cRIK4Xavpbbed7yg7Cy59p0a3lZPwteTBP&uniplatform=NZKPT&language=CHS
- Wen, Y. et al. A data-driven method of traffic emissions mapping with land use random forest models. *Appl. Energy* **305**, 117916 (2022).
- Zhang, J. et al. Measuring the critical influence factors for predicting carbon dioxide emissions of expanding megacities by XGBoost. *Atmosphere* **13** (4), 599 (2022).
- Luo, H. et al. Unveiling land use-carbon nexus: Spatial matrix-enhanced neural network for predicting commercial and residential carbon emissions. *Energy* **305**, 131722 (2024).
- Wen, L. & Yuan, X. Forecasting CO₂ emissions in Chinas commercial department, through BP neural network based on random forest and PSO. *Sci. Total Environ.* **718**, 137194 (2020).
- He, Z. et al. Path to Peak Emissions and Smart Urbanization. *Urban Planning Forum*(06), 37–44. Retrieved from <https://link.cnki.net/doi/> (2021). <https://doi.org/10.16361/j.upf.202106005> doi:10.16361/j.upf.202106005
- Huang, Y., Li, S., Lin, J., Zheng, L., Zhuang, C., Guan, C., ... Zhuang, Y. (2025).Nonlinear and threshold effects of urban building form on carbon emissions. *Energy and Buildings*, 329, 115243.
- Wang, Q. & Chen, X. Energy policies for managing china's carbon emission. *Renew. Sustain. Energy Rev.* **50**, 470–479 (2015).
- Agency, I. E. *Cities, Towns & Renewable Energy: Yes in my Front Yard* (OECD/IEA, 2009).

30. Kwok, K., Kim, J., Chong, W. K. & Ariaratnam, S. T. Structuring a comprehensive carbon-emission framework for the whole lifecycle of building, operation, and construction. *J. Archit. Eng.* **22** (3), 04016006 (2016).
31. Ramesh, T., Prakash, R. & Shukla, K. K. Life cycle energy analysis of buildings: An overview. *Energy and Buildings*, **42**(10), 1592–1600. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0378778810001696>. (2010). <https://doi.org/10.1016/j.enbuild.2010.05.007>
32. Baek, C., Park, S. H., Suzuki, M. & Lee, S. H. Life cycle carbon dioxide assessment tool for buildings in the schematic design phase. *Energy Build.* **61**, 275–287 (2013).
33. Gardezi, S. & Shafiq, N. Operational carbon footprint prediction model for conventional tropical housing: a Malaysian prospective. *Int. J. Environ. Sci. Technol.* **16** (12), 7817–7826 (2019).
34. Ewing, R. & Cervero, R. Travel and the built environment: A meta-analysis. *J. Am. Plann. Assoc.* **76** (3), 265–294 (2010).
35. Yu, T., Leng, H., Yuan, Q., LOW-CARBON SPATIAL & PLANNING TECHNOLOGY IN THEURBAN AREA OF COUNTIES UNDER THE CARBON PEAKING AND CARBONNEUTRALITY GOALS. A STUDY ON. *City Planning Review*, **47**(06), 110–120. (2023). Retrieved from <https://link.cnki.net/urlid/11.2378.TU.20230421.1656.006>
36. Depecker, P., Ménézo, C., Virgone, J. & Lepers, S. Design of buildings shape and energetic consumption. *Build. Environ.* **36**, 627–635. [https://doi.org/10.1016/S0360-1323\(00\)00044-5](https://doi.org/10.1016/S0360-1323(00)00044-5) (2001).
37. Alhorri, Y., Eliskandarani, E. & Elsarrag, E. Approaches to reducing carbon dioxide emissions in the built environment: low carbon cities. *Int. J. Sustainable Built Environ.* **3** (2), 167–178 (2014).
38. Filogamo, L., Peri, G., Rizzo, G. & Giaccone, A. On the classification of large residential buildings stocks by sample typologies for energy planning purposes. *Appl. Energy*, **135**, 825–835 (2014).
39. Shen, Y. S., Lin, Y. C., Cui, S., Li, Y. & Zhai, X. Crucial factors of the built environment for mitigating carbon emissions. *Sci. Total Environ.* **806**, 150864 (2022).
40. Sheugh, L. & Alizadeh, S. H. A note on pearson correlation coefficient as a metric of similarity in recommender system. Paper presented at the 2015 AI & Robotics (IRANOPEN). (2015).
41. Chen, X., Shuai, C., Wu, Y. & Zhang, Y. Analysis on the carbon emission peaks of china's industrial, building, transport, and agricultural sectors. *Sci. Total Environ.* **709**, 135768 (2020).
42. Wei, D., Hu, X., Chen, Y., Li, B. & Chen, H. An investigation of the quantitative correlation between urban Spatial morphology indicators and block wind environment. *Atmosphere* **12** (2), 234 (2021).
43. Faroughi, M. et al. Computational modeling of land surface temperature using remote sensing data to investigate the Spatial arrangement of buildings and energy consumption relationship. *Eng. Appl. Comput. Fluid Mech.* **14** (1), 254–270 (2020).
44. Ryu, H., Park, I. K., Chun, B. S. & Chang, S. I. Spatial statistical analysis of the effects of urban form indicators on road-traffic noise exposure of a City in South Korea. *Appl. Acoust.* **115**, 93–100 (2017).
45. Gan, V. J., Chan, C. M., Tse, K., Lo, I. M. & Cheng, J. C. A comparative analysis of embodied carbon in high-rise buildings regarding different design parameters. *J. Clean. Prod.* **161**, 663–675 (2017).
46. Su, X. & Zhang, X. A detailed analysis of the embodied energy and carbon emissions of steel-construction residential buildings in China. *Energy Build.* **119**, 323–330 (2016).
47. Victoria, M. F. & Perera, S. Parametric embodied carbon prediction model for early stage estimating. *Energy Build.* **168**, 106–119 (2018).
48. Xikai, M., Lixiong, W., Jiwei, L., Xiaoli, Q. & Tongyao, W. 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. *Energy Build.* **204**, 109519 (2019).
49. Gardezi, S. S. S., Shafiq, N., Zawawi, N. A. W. A., Khamidi, M. F. & Farhan, S. A. A multivariable regression tool for embodied carbon footprint prediction in housing habitat. *Habitat Int.* **53**, 292–300 (2016).
50. Liu, L. et al. Effect of geometric factors on the energy performance of high-rise office towers in Tianjin, China. Paper presented at the Building Simulation. (2017).
51. Rong, P., Zhang, Y., Qin, Y., Liu, G. & Liu, R. Spatial differentiation of carbon emissions from residential energy consumption: A case study in kaifeng, China. *J. Environ. Manage.* **271**, 110895 (2020).
52. Zhang, X., Yan, F., Liu, H. & Qiao, Z. Towards low carbon cities: A machine learning method for predicting urban blocks carbon emissions (UBCE) based on built environment factors (BEF) in Changxing city, China. *Sustainable Cities Soc.* **69**, 102875 (2021).
53. Oh, M., Jang, K. M. & Kim, Y. Empirical analysis of Building energy consumption and urban form in a large city: A case of seoul, South Korea. *Energy Build.* **245**, 111046 (2021).
54. Basbagill, J., Flager, F., Lepech, M. & Fischer, M. Application of life-cycle assessment to early stage Building design for reduced embodied environmental impacts. *Build. Environ.* **60**, 81–92 (2013).
55. Lin, J., Lu, S., He, X. & Wang, F. Analyzing the impact of three-dimensional Building structure on CO2 emissions based on random forest regression. *Energy* **236**, 121502 (2021).
56. Motulsky, H. J. & Ransnas, L. A. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. *FASEB J.* **1** (5), 365–374 (1987).
57. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
58. Genuer, R., Poggi, J. M. & Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **31** (14), 2225–2236 (2010).
59. Ye, T. et al. Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. *Sci. Total Environ.* **658**, 936–946 (2019).
60. Yoon, J. Forecasting of real GDP growth using machine learning models: gradient boosting and random forest approach. *Comput. Econ.* **57** (1), 247–265 (2021).
61. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. (2016).
62. Deng, J. Y. & Wong, N. H. Impact of urban Canyon geometries on outdoor thermal comfort in central business districts. *Sustainable Cities Soc.* **53**, 101966 (2020).
63. Timmons, D., Zirogiannis, N. & Lutz, M. Location matters: population density and carbon emissions from residential Building energy use in the united States. *Energy Res. Social Sci.* **22**, 137–146 (2016).
64. Ewing, R. & Rong, F. The impact of urban form on US residential energy use. *Hous. Policy Debate*, **19** (1), 1–30 (2008).
65. Kohansal, M. E., Akaf, H. R., Gholami, J. & Moshari, S. Investigating the simultaneous effects of Building orientation and thermal insulation on heating and cooling loads in different climate zones. *Architectural Eng. Des. Manage.* **18** (4), 410–433 (2022).
66. Antonopoulos, C., Trusty, A. & Shandas, V. The role of Building characteristics, demographics, and urban heat Islands in shaping residential energy use. *City Environ. Interact.* **3**, 100021 (2019).
67. Liu, S., Wang, J., Xu, Y. & Zeng, H. Approaches of buildings carbon mitigation based on Spatial-Temporal characteristics of civil Building energy consumption: A case study on shenzhen, China. *Acta Scientiarum Naturalium Universitatis Pekinensis*, **54** (01), 125–136 (2018).
68. Feng, G., Sha, S. & Xu, X. Analysis of the Building envelope influence to Building energy consumption in the cold regions. *Procedia Eng.* **146**, 244–250 (2016).
69. Sozer, H. Improving energy efficiency through the design of the Building envelope. *Build. Environ.* **45** (12), 2581–2593 (2010).
70. Djuric, N., Novakovic, V., Holst, J. & Mitrovic, Z. Optimization of energy consumption in buildings with hydronic heating systems considering thermal comfort by use of computer-based tools. *Energy Build.* **39** (4), 471–477 (2007).

71. Yun, G. Y., Kim, H. & Kim, J. T. Effects of occupancy and lighting use patterns on lighting energy consumption. *Energy Build.* **46**, 152–158 (2012).

Author contributions

Formal analysis, writing paper, software and original draft preparation, Y.X.; conceptualization, methodology, reviewing and editing, supervision, X.C; visualization and investigation L.S, and T.Y. All authors reviewed the manuscript.

Funding

This study was supported by the National Natural Science Foundation of China (No. U20A20330).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-17746-z>.

Correspondence and requests for materials should be addressed to Y.X. or L.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025