Quantifying the heterogeneous impacts of the urban built environment on traffic carbon emissions
New insights from machine learning techniques

Zhi, Danyue; Zhao, Hepeng; Chen, Yan; Song, Weize; Song, Dongdong; Yang, Yitao

DOI
10.1016/j.uclim.2023.101765

Publication date
2024

Document Version
Final published version

Published in
Urban Climate

Citation (APA)
Zhi, D., Zhao, H., Chen, Y., Song, W., Song, D., & Yang, Y. (2024). Quantifying the heterogeneous impacts of the urban built environment on traffic carbon emissions: New insights from machine learning techniques. *Urban Climate*, *53*, Article 101765. https://doi.org/10.1016/j.uclim.2023.101765

Urban Climate 53 (2024) 101765

Contents lists available at ScienceDirect

Urban Climate

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

Quantifying the heterogeneous impacts of the urban built environment on traffic carbon emissions: New insights from machine learning techniques

Danyue Zhi [a,b], Hepeng Zhao [c], Yan Chen [d], Weize Song [e,f,*], Dongdong Song [a,*], Yitao Yang [a,g,*]

[a] *School of Systems Science, Beijing Jiaotong University, Beijing 100044, China*
[b] *TUM School of Engineering and Design, Technical University of Munich, Munich 80333, Germany*
[c] *China Mobile Group Design Institute Co., Ltd., Beijing 100080, China*
[d] *School of Traffic and Transportation, Beijing Jiaotong University, 100044, China*
[e] *Laboratory of Low Carbon Energy, Tsinghua University, Beijing 100084, China*
[f] *Institute of Climate Change and Sustainable Development, Tsinghua University, Beijing 100084, China*
[g] *Department of Transport & Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg1, Delft 2628, CN, the Netherlands*

A R T I C L E I N F O

Keywords:
Traffic carbon
Urban built environment
Machine learning
Heterogeneous impact

A B S T R A C T

The configuration of the urban built environment is critical for promoting sustainability and achieving carbon neutrality. However, existing studies mostly use linear and spatial econometric models to investigate the relationship between urban built environments and traffic carbon dioxide (CO_2) emissions, in-depth studies exploring the heterogeneous impacts of related features on traffic CO_2 emission by interpretive machine learning models are scarce. Hence, we extract four dimensionless features to depict the size, compactness, irregularity, and isolation of built-up areas, and road network-related features (i.e., average cluster coefficient, road topological density, and road geometric density), respectively. Subsequently, we develop an interpretive machine learning framework based on the extracted features related to the urban built-up areas and road networks. The interpretive results of the proposed framework uncover that urban morphological features, especially population density (POP), GDP per capita (GDPpc), and urban physical compactness (UPC), have a heterogeneous impact on the per capita traffic emission (PCCE) across different cities. GDPpc is more like a linear relationship with PCCE, and UPC has a significant influence on PCCE when its value is between 62% and 78%. Our results also reveal the nonlinear relationships and interactive effects between these features, providing the implications of urban morphological planning and carbon emission reduction.

1. Introduction

The carbon emissions of urban areas emit >70% worldwide (Axsen et al., 2020). Urban transportation is the sector with the fastest-growing emissions, and it ranks second among all sectors with regard to the consumption of energy and emissions of carbon dioxide

* Corresponding authors.
E-mail addresses: songwz@mail.tsinghua.edu.cn (W. Song), dongdongsong@bjtu.edu.cn (D. Song), yitao-yang@bjtu.edu.cn (Y. Yang).

https://doi.org/10.1016/j.uclim.2023.101765
Received 7 March 2023; Received in revised form 28 October 2023; Accepted 16 November 2023
Available online 24 November 2023
2212-0955/

(CO_2) with 29% and 24%, respectively (IEA, 2018). In the future, urban transport carbon emissions are rising rapidly as the urban space expands and car ownership remains growing, becoming an important issue of current urban governance (Shi et al., 2022). The configuration of the urban built environment includes features such as the design and layout of streets, buildings, parks, and public spaces, as well as the transportation infrastructure that connects them, affecting how people travel and the amount of carbon emitted in the process. Quantifying the heterogeneous impacts of the urban built environment on traffic CO_2 emission is critical for promoting sustainability, and livability and achieving carbon neutrality.

Many studies have shown that changes in the built environment are closely related to changes in urban traffic CO_2 emission, and this is a complex process with multiple effects (Kwak et al., 2018; Huang et al., 2022). In built-up areas, scale/size refers to the degree of expansion of the urban area: with the development of urbanization, cultivated land, forests, and grasslands are continuously transformed into construction land. Numerous studies have indicated that preserving green spaces contributes to the augmentation of CO_2 carbon sinks, while the expansion of developed land leads to an increase in CO_2 emissions (Liu et al., 2014; Xu et al., 2016; Salvia et al., 2021). Built-up areas' irregularity and compactness also affect CO_2 emissions at multiple spatial scales (Liu et al., 2017b; Guan et al., 2019; Song et al., 2021). Moreover, Yang et al. (2020) and Mohajeri et al. (2015) analyzed the impacts of the street network (length entropy, average road length, street density) on traffic CO_2 emissions. Barrington-Leigh and Millard-Ball (2017) found that more connected roads effectively make contributions to CO_2 emission mitigation. Other studies also suggested that population density, and per capita GDP are the main features that affect traffic CO_2 emissions (Wang et al., 2018; Wang et al., 2019).

Previous studies have devoted a great deal of effort to identifying the impact of the urban built environment on traffic CO_2 emissions by using various approaches, such as the structural decomposition analysis (SDA) (Hoekstra and van der Bergh, 2003), exponential decomposition analysis (IDA) (Zhang et al., 2009), data envelopment analysis (DEA) (Ma et al., 2022), Stochastic Impacts by Regression on Population, Affluence, and Technology (STIRPAT) (Liu et al., 2017a), geographically Weighted Regression (GWR) (Zhou et al., 2022b), and Spatial autoregressive model (SAM) (Song et al., 2021). However, the effects of the built environment attributes on the activity space and mode of transportation use are complex, such as nonlinear relationships and threshold effects, and the traditional linear regression model has a low fit (Tu et al., 2021; Xu et al., 2021; Duan et al., 2023). To the best of our knowledge, the complex nonlinear relationship between the built environment and traffic CO_2 emissions in multiple cities has not been revealed. The advent of the big data era provides the possibility to address this issue from a new perspective of machine learning. Machine learning models have been widely used in a variety of fields in recent years. (Ou et al., 2020; Yu et al., 2020; Lv et al., 2021; Park et al., 2022). Compared to traditional linear models and spatial econometric models, machine learning models can handle nonlinear interactions between features. In addition, machine learning models generally have higher potential and advantages in the accuracy of fitting data, especially for complex non-linear relationships (Tu et al., 2021; Xu et al., 2021). Chen and Guestrin are the first to propose XGBoost (Chen and Guestrin, 2016), and its predictive performance is superior to many other machine learning methods, such as K Nearest Neighbors and Gradient Boosted Decision Trees, etc. (Wang et al., 2021). Although machine learning models can be black boxes, researchers have developed techniques to explain the model's decision-making process. These techniques, known as "interpretability" approaches, aim to provide insight into how models make decisions. For example, SHAP (Shapley Additive exPlanation), a widely used interpretive model, was initially proposed by Shapley and has been used to explain the results of machine learning models (Parsa et al., 2020; Dikshit and Pradhan, 2021; Lei et al., 2022; Park et al., 2022).

To sum up, previous studies on the impact of the built environment on traffic CO_2 lacked an integrated consideration of built-up areas and road networks, and there was a lack of extracting the relevant density and connectivity characteristics of road networks from the perspective of complex networks. Moreover, most previous studies are based on traditional linear and spatial econometric models, which are unable to adequately capture the complexity and nonlinear of the real-world urban built environment and lack new perspectives and insights from machine learning models.

This study aims to develop an interpretable machine learning framework to capture the heterogeneous influence of the urban built environment on traffic CO_2 emission. We characterize the configurations of the urban built environment from the perspectives of urban morphology (i.e., built-up area) and road networks. The three specific objectives are as follows: (1) what is the relative importance of urban morphology and road networks on traffic CO_2 emission, and their heterogeneous impacts; (2) what are the nonlinear relationships between traffic CO_2 emission and the configurations of the urban built environment; (3) what are the potential interactions between the multiple features characterized the configurations of the urban built environment.

In summary, the main contributions of this study include the following:

(1) This study applies interpretable machine learning models (XGBoost, SHAP, and Accumulated local effects (ALE)) to explore traffic CO_2 emissions, revealing the nonlinear relationship between influencing features and CO_2 emissions.
(2) This study provides insights into the influence of the built-up area and road network features on traffic CO_2 from both overall and individual perspectives, and reveals the importance and heterogeneous impact of these features.
(3) Based on the machine learning explanatory models (i.e., SHAP and ALE), this study further studies the interactive effects of two groups of features on traffic CO_2 emissions.

The rest of the study is organized as follows. Section 2 is the literature related to traffic CO_2. Section 3 shows the methodology. Section 4 analyzes the results of the machine learning model. Section 5 summarizes the conclusion and discussions.

2. Related works

2.1. Characteristics analysis of traffic CO_2 emissions

Most studies focus on the spatial and temporal characteristics of traffic CO_2 emissions, especially the spatial characteristics (Zhou and Li, 2014; Yang et al., 2015; Bai et al., 2019; Li et al., 2019; Liu et al., 2022; Ma et al., 2022). In particular, Yang et al. (2015) used a transportation carbon-emission model to examine the spatiotemporal trends of traffic CO_2 emissions from 2000 to 2012 in China. Studying the spatial characteristics of traffic CO_2 emissions mainly focuses on spatial imbalance, spatial heterogeneity, spatial spillover effect, spatial evolution pattern, and spatial association network structure (Bai et al., 2019; Song et al., 2021; Ma et al., 2022). For instance, Bai et al. (2019) used a social network analysis model to investigate the characteristics of the geographical association network structure of province-level traffic CO_2. In order to identify the spatial imbalance, Ma et al. (2022) assessed the transport carbon emission efficiency (TCEE) of Chinese provinces from 2003 to 2018. The findings indicate that the TCEE is rising at a generally modest rate with notable regional variations; the regional variations of TCEE exhibit a tendency to rise initially and then decline, with intra-regional variations constituting the major source of total variations.

2.2. Influencing features of traffic CO_2 emissions

It is crucial to examine the motivating or determining reasons behind changes in socio-economic indicators like the economy, environment, and employment in order to examine and comprehend those changes. At present, there are two methods mainly used for the quantitative decomposition of indicators: SDA and IDA. Among these decomposition methods, there are two typical methods according to the principle of index decomposition: Laspeyres index decomposition and Divisia index decomposition. Research on traffic CO_2 dioxide is mostly focused on the country level, provincial level, and city level, as summarized in Table 1. In the table, the LDMI model was applied to explore the important features of traffic CO_2 literature at the country level (Zhu and Du, 2019; Liu et al., 2021). LDMI is a Divisia index decomposition method in IDA, which has no unexplained residuals after the object is decomposed, and can use addition decomposition and multiplication decomposition for relatively simple transformation expressions. In addition, the spatial econometric models are also applied to examine the influencing features, which mainly focus on the influence of space (Lv et al., 2019; Liu et al., 2022). In particular, Liu et al. (2022) investigated the driving force behind emissions decrease and the effect of technological advancement on spillovers of CO_2. The findings demonstrated that the "community effect" caused interactions between traffic CO_2 reduction practices in different provinces, and advancements in transportation technology were more likely to happen between regions that had seen similar transportation growth. It is worth noting that with the narrowing of the research scope, the

Table 1
Related works about the influencing features of traffic CO_2 emission.

References	Scale	Study object	Country (city)	Method	Key influencing factor
Zhu and Du (2019)	Country Level	CO_2 of the road transportation industry	Australia, Canada, China, India, Russia, and the United States	LMDI	Economic output; population size; energy intensity; transportation intensity
Liu et al. (2021)		CO_2 of the transport sector and the effect of economic growth	China	LMDI	Energy emission intensity; energy structure; transportation intensity;
Bai et al. (2019)	Provinces Level	Traffic CO_2 emissions	China	LMDI	Transportation energy intensity; per-capita wealth
Lin et al. (2019)		Process of traffic CO_2 emissions	China	Panel model.	Socioeconomic features; regional variations
Lv et al. (2019)		Impact of urbanization on traffic CO_2	China	SDM	Urbanization
Liu et al. (2022)		Impact of technological progress on regional CO_2 emissions spillovers	China	SDM	Technological progress
Ma et al. (2022)		Regional differences of TCEE	China	DEA	Economic and technical level; transport and energy structure;
Rong et al. (2018)		CO_2 arising from the travel activities of residents	Kaifeng, China	GWR	Urban sprawl
Yang et al. (2020)		Traffic CO_2 emissions in fast-growing cities	China, India	Transport CO_2 emission model	Per capita GDP
Song et al. (2021)		Traffic CO_2 emissions	China	SAM	Built environment (built-up area's size, compactness, irregularity, isolation)
Zhou et al. (2022b)		Road traffic CO_2 emissions	Shenzhen, China	GWR	Built environment (density of the main road, access to metro stations/bus stops)
Zhang et al. (2023)		Traffic CO_2 emissions	19 cities, China	Principal component analysis	Built environment (road network form)
This study	City Level	Heterogeneous and nonlinear impact of the urban built environment on traffic CO_2 emission	China	• XGBoost; • SHAP; • ALE	• Built-up area features • Road network features • Per capita GDP

influencing features selected are becoming more and more specific and detailed.

Traffic CO_2 emissions at the country level are largely influenced by energy and economy-related factors, such as economic output, energy intensity, and energy structure. As the research scope changes to the province level, the impact of urbanization-related factors on traffic CO_2 on the basis of economic energy has also been verified. Specifically, Lv et al. (2019) found urbanization level has a significant positive impact on road and air transport CO_2 emissions in some provinces and a significant negative impact on rail and water transport CO_2. It is worth noting that relevant studies at the city level have focused more on the impact of the built environment on transport CO_2 (Wang et al., 2017; Zhang et al., 2018; Wu et al., 2019; Song et al., 2021; Ashik et al., 2023). The study of Ashik et al. (2023) showed the built environment played a different role in explaining CO_2 emissions from commuting in developed and developing countries. Song et al. (2021) adopted landscape metrics (i.e., built-up areas' size, compactness, irregularity, and isolation) to quantify the built environment configuration differences between cities. The findings showed that a 1% increase in built-up areas' physical size, compactness, and isolation are associated with increases of 0.35%, −0.14%, and 0.13%, respectively, in traffic CO_2 emissions in adjacent cities. Furthermore, many studies have considered the effect of urban road network characteristics in the built environment on traffic CO_2 (Sun et al., 2017; Wang et al., 2017; Chen, 2023; Zhang et al., 2023). The study of Wang et al. (2017) indicated that transportation factors, such as urban road density and traffic coupling, had significant negative effects on CO_2 emissions. Zhang et al. (2023) found that there were significant differences between the carbon emissions of different road network forms and road network form transformation could reduce annual traffic CO_2 emissions by 1.13% to 14.35%.

However, in the study of CO_2 emissions from city-level transportation, there is a lack of integrated consideration of the built-up area and road network, as well as a lack of selection of relevant features of the urban road network based on the complex network methodology. In addition, the previous studies mostly used traditional decomposition models or spatial econometric models.

2.3. Application of machine learning models to sustainable city

Machine learning has been extensively used in various domains as science and technology have advanced. This section briefly introduces its application in sustainable city research. Machine learning models have started to be used in studies on CO_2 emissions and urban development in recent years (Li et al., 2022). Different from the various linear regression models and spatial econometric methods, machine learning models can explore a more complex relationship between urban characteristics and the urban environment, such as CO_2 emission, ozone, urban heat island, urban stormwater, heat-related mortality, indoor water leakage (Kourtit et al.,

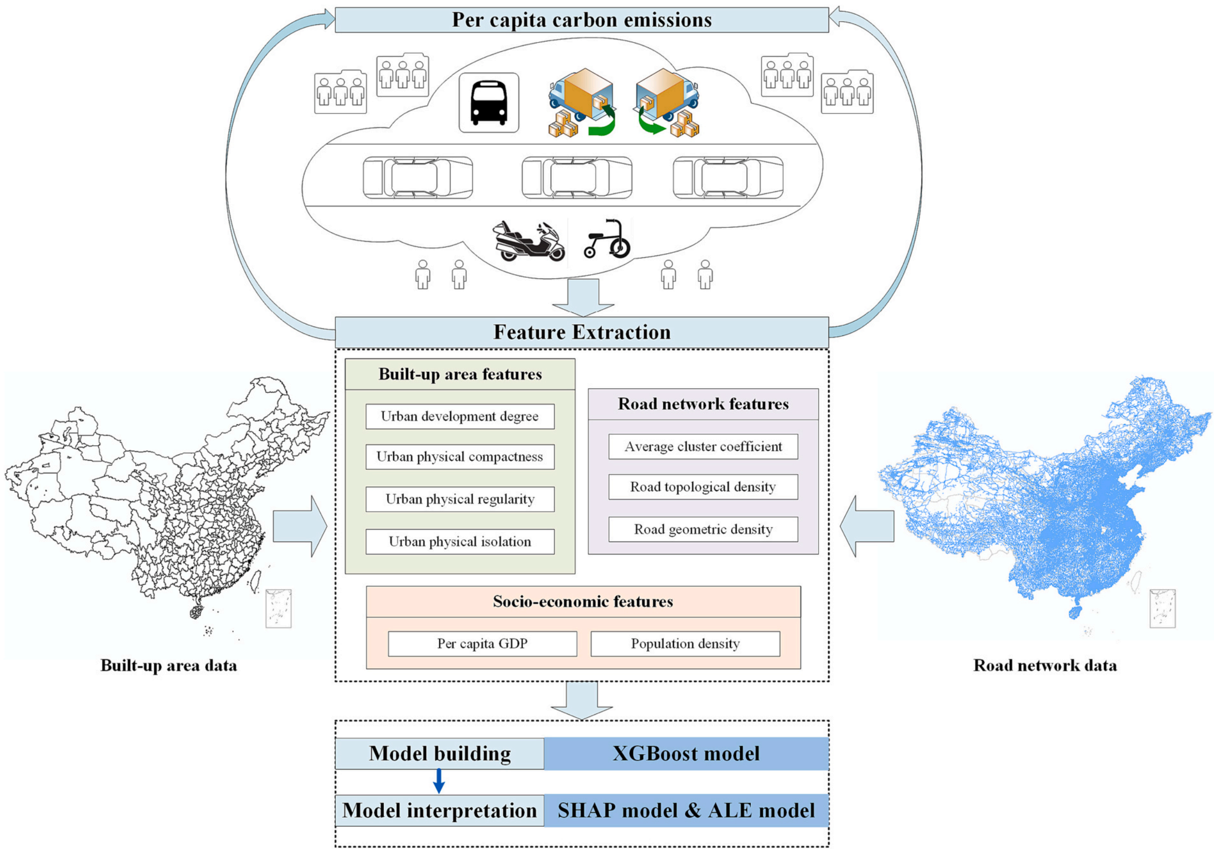

Fig. 1. The study framework layout.

2021; Kim and Kim, 2022; Han et al., 2022; Haase, 2022; Shin et al., 2022; Zhou et al., 2022a). Excess anthropogenic carbon emissions as a result of fast socio-economic growth and urbanization caused severe global warming. Combining XGBoost and multiscale GWR models, Li et al. (2022) investigated the spatial heterogeneity of socioeconomic and city-shape features on CO_2 emissions in different cities. They found that, whereas urban forms had less impact on CO_2 than per capita GDP, the contribution from urban forms rose quickly.

However, the complex nonlinear relationship between traffic CO_2 emissions and the built environment combining built-up area and road network features is not revealed at present. Meanwhile, the study of the interaction of these built environment features on traffic CO_2 emissions has not been investigated, which can help to determine the synergistic and antagonistic effects between different features. Machine learning methods can model the complex relationship between features and traffic CO_2 and show high accuracy over traditional models, which can lead to more insightful results.

3. Methodology

The spatial analysis is carried out with 275 Chinese cities as samples, i.e., the study unit is the city administrative area. Data on traffic CO_2 emissions are extracted from the city-level CO_2 emission inventory of China High-Resolution Emission Database. Our data provide information on CO_2 emissions of passenger and freight vehicles, including motorcycles, low-speed freight vehicles, and three-wheeled vehicles.

In this study, we first extract the features based on built-up area and road network to represent the characteristics of the studied unit. Then use the XGBoost model to construct the relationships between these extracted features and traffic CO_2 emissions. We use the SHAP and ALE models to interpret the results. We aim to find the features affecting traffic CO_2 emissions and explore their complex relationships and interactions. The modeling framework is displayed in Fig. 1. In detail, we utilize the following Python packages and environments: XGBoost Version 1.2.0, SHAP Version 0.36.0, PyALE Version 1.1.3, and Anaconda Version 4.14.0. The computational tasks are performed on a system equipped with an Intel Core i7-9700 K CPU with 8 cores, 16 GB of DDR4 RAM operating at 3200 MHz, and running the Windows 10 Pro 64-bit operating system.

3.1. *Feature extraction*

The four widely recognized elements of transportation are people, vehicles, roads, and the environment, which affect traffic CO_2 emissions (Yang, 2003). The variables selected in this study are those related to the four elements, especially the "environment". Specifically, population-related features and economy-related GDP are often associated with travel demand and urban car ownership (Ma et al., 2019). Road network features reflect road infrastructure conditions, which affect the operating efficiency of the traffic system and thus CO_2 emissions (Barrington-Leigh and Millard-Ball, 2017; Yang et al., 2020). Furthermore, attributes of urban built-up areas can affect urban construction, land use, and road planning, which, in turn, can influence travel behavior and CO_2 emissions (Guan et al., 2019; Song et al., 2021).

The features of the built-up area, such as scale, compactness, and irregularity, play a crucial role in traffic CO_2 emissions (Liu et al., 2017b; Wang et al., 2017; Guan et al., 2019; Song et al., 2021). Additionally, research findings have indicated that features of the road network, including road density and connectivity, also contribute to CO_2 emission mitigation (Mohajeri et al., 2015; Barrington-Leigh and Millard-Ball, 2017; Wang et al., 2017; Yang et al., 2020). The inclusion of built-up area features and road network features provides vital information about the urban transportation system and its impact on CO_2 emissions. Therefore, this study extracts features related to size, compactness, irregularity, isolation of built-up areas, and network density and connectivity of the road network, which are from the urban dataset and the road network dataset.

Table 2
The statistics of features.

Feature	Mean	S.D.	Min	Max
PCCE	548.970273	321.452741	135.865805	2202.225617
Socio-economic features				
GDPpc	8262.446423	4938.612842	1111.950014	33,719.801394
POP	440.822073	340.882904	5.770000	2501.140000
Built-up area features				
UDD	2.170855	4.200459	0.027040	41.596802
UPC	70.866844	13.328039	34.178405	96.775320
UPR	30.269053	3.982616	13.222197	37.048616
UPI	0.206009	0.372717	0.002741	3.857456
Road network features				
ACC	0.001010	0.001114	0.000000	0.006494
RTD	0.000703	0.000508	0.000060	0.003375
RGD	0.589642	0.464227	0.046014	3.711563

3.1.1. Built-up area features

Landscape indices are an objective and standardized measurement method that comprehensively considers multiple features of the built-up area, providing rich information about the structure and organization of the built environment (Liu et al., 2017b; Guan et al., 2019). Specifically, landscape indices generally measure the size, compactness, irregularity, and isolation of built-up areas. *Size* refers to the geographical coverage of urban patches, which can have implications for factors such as transportation efficiency, access to services and amenities, and the availability of open space. *Compactness* refers to the degree to which buildings and other structures are clustered together, as opposed to being spread out over a large area. More compact urban areas tend to be more walkable and bikeable and can support a wider range of public transportation options. *Irregularity* refers to the shape and arrangement of urban patches, which can have implications for issues such as accessibility, and land use efficiency. *Isolation* refers to the degree to which the built-up area is physically separated from other urban areas, either by natural barriers such as water bodies or by man-made barriers such as highways or industrial zones.

The built-up area dataset in our study is obtained from the Chinese Academy of Science Data Centre for Resources and Environmental Sciences (Chinese Academy of Sciences, et al., 2020). Based on the original landscape indices, the dimensionless urban development degree (UDD), urban physical compactness (UPC), urban physical regularity (UPR), and urban physical isolation (UPI) are proposed to depict the size, compactness, irregularity, and isolation of built-up areas in this study. It should be noted that the built-up area of a city is formed by a series of land parcels. The calculation of these built-up area features is shown below (Song et al., 2021). Table 2 summarizes the statistical values of these built-up area features.

3.1.1.1 Urban development degree (UDD)

$$UDD = \sum\nolimits_{j=1}^{n} a_j \Big/ A \times 100 \tag{1}$$

where a_j represent the area of land parcel j (unit: m^2). n shows the sum of the number of urban land parcels. A is the total area of the urban district (unit: m^2).

3.1.1.2 Urban physical compactness (UPC)

$$UPC = \left(1 - \frac{\sum_{j=1}^{n} l_j}{\sum_{j=1}^{n} l_j \sqrt{a_j}}\right) \cdot \left(1 - \frac{1}{\sqrt{n}}\right)^{-1} \times 100 \tag{2}$$

where l_j is the perimeter of land parcel j. a_j have the same meanings as the variables in Eq. (1).

3.1.1.3 Urban physical regularity (UPR)

$$UPR = \frac{1}{n} \sum_{j=1}^{n} \frac{l_j}{a_j} \tag{3}$$

where l_j,a_j, and n have the same meanings as the variables in Eq. (1).

3.1.1.4 Urban physical isolation (UPI)

$$UPI = \frac{1}{n} \sum_{j=1}^{n} \min_{i \neq j} d_{ij} \tag{4}$$

where $d_{i,j}$ is the distance between land parcels i and j. n have the same meanings as the variables in Eq. (1).

3.1.2. Road network features

In this study, the connectivity and density of the actual road network in the study unit are calculated based on complex network theory (Scalfani et al., 2022; Wei et al., 2022). The actual road network data of Chinese cities are obtained from the OpenStreetMap, consisting of five road types, i.e., high-speed highways, highways, primary, secondary, and tertiary roads. The starting/ending point of each road section is retrieved using the Python package OSMnx (Boeing, 2017). In this study, based on the extracted road network of each city, the average cluster coefficient (ACC) is selected to evaluate the level of clustering (i.e., connectivity) in road networks of different cities. Networks with a high ACC indicate that road intersections tend to be highly interconnected, while networks with a low ACC indicate that road intersections are less likely to be interconnected. In addition, road topological Density (RTD) and road geometric density (RGD) are used to measure road density from the perspective of a complex network and an actual road network, respectively. It is worth noting that RTD is the ratio of the number of actual edges in the network to the number of all possible edges. Urban road network features are calculated as follows and Table 2 has a summary of the statistical values of these road network features.

3.1.2.1 Average cluster coefficient (ACC)

$$ACC = \frac{1}{n} \sum_{i=1}^{n} \sum\nolimits_{v \in N(i)} \frac{|N(i) \cap N(v)|}{|N(i)^*(N(i) - 1)|} \tag{5}$$

where i is the vertex in the network (i.e., the urban road network in the study). $N(i)$ is the set of neighbors of vertex i. n represents the total number of road nodes. This metric measures the degree of node agglomeration, which reflects the closeness of the whole network.

3.1.2.2 Road topological Density (RTD)

$$RTD = \frac{2E}{n(n-1)} \tag{6}$$

where n is the number of nodes and $n(n-1)$ represents that there can be at most $n(n-1)$ undirected edges when at most one edge between each pair of nodes. E is the number of edges in the road network, i.e., the number of actually existing edges.

3.1.2.3 Road geometric density (RGD)

$$RGD = \frac{L}{A} \tag{7}$$

where L is the total length of roads. A is the total area of the urban district. In addition to the above two types of features, GDPpc and POP are particularly used to characterize the social and economic status at the city level, as shown in the following.

3.1.2.4 Per capita GDP (GDPpc) (unit: dollar per capita)

$$GDP_{pc} = \frac{GDP}{P} \tag{8}$$

where GDP is the (unit: ten thousand dollars). P is the population of the city (unit: ten thousand people). Population density POP is obtained from the Statistical Yearbook of Chinese Cities (Unit: people/km^2).

3.1.2.5 Per capita carbon emissions (PCCE) (unit: kg per capita)

Finally, we choose the per capita carbon emissions (PCCE) to be the explained variable of models, calculated as follows:

$$PCCE = \frac{CE}{P} \times 1000 \tag{9}$$

where CE is the urban traffic CO_2 emission (unit: ten thousand tonnes). P is the population of the city (unit: ten thousand people) as shown in Eq. (8).

CE in Eq. (9) is obtained from the Chinese city-level CO_2 emission inventory, which is published in the China High Resolution Emission Database (City Greenhouse Gas, 2020). This inventory accounts for the overall CO_2 emissions originating from a range of road vehicles, encompassing passenger vehicles, freight vehicles, three-wheeled vehicles, low-speed freight vehicles, and motorcycles (Song et al., 2021). The CO_2 emissions in this inventory are calculated based on data about gasoline and diesel consumption using a bottom-up method. For more comprehensive information, please consult the detailed study conducted by Cai et al. (2019).

3.2. Modeling and interpretation

In the above, we extract the built-up area and road network features of each city. Next, we use the XGBoost model to construct the relationships between these features and PCCE, and then use SHAP and ALE models to interpret the results of XGBoost.

3.2.1. Machine learning prediction model: XGBoost

The XGBoost model of Chen and Guestrin (2016) was proposed based on the original gradient-boosting framework. Given a dataset with n samples (in this study n is taken as 275 cities), there are independent features x_i of the city i (i.e., 9 features extracted in Section 3.1), and each city has m features (i.e., $x_i \in R^m$, in this study m is taken as 9). Each feature has an associated dependent variabley_i ($y_i \in R$), i.e., PCEE in this study. Finally, the tree ensemble model is applied to predict the dependent variable, which is $\hat{y}_i$ in Eq. (10) (Chen and Guestrin, 2016):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{10}$$

wheref_k is the structure of individual trees with leaf scores, and F is the range of f_k.

The goal of the XGBoost model is to minimize $\mathscr{L}(\phi)$ in Eq. (11):

$$\mathscr{L}(\phi) = \sum_i L(\hat{y}_i, y_i) + \sum_k \mathrm{M}(f_k) \tag{11}$$

where L is the function of precision loss. M is the penalty function for model complexity.

$$\mathrm{M}(f) = \gamma l + \frac{1}{2} \|S_i\|^2 \tag{12}$$

where l is the number of leaves. S_i denotes the score of the ith leaf.

After using GBDT to represent the XGBoost formula, the objective function is shown below after the second-order Taylor expansion and regularized term expansion and combining the coefficients.

$$\tilde{\mathscr{L}}^{(t)} = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\omega_j^{\,2}\right] + \gamma T \tag{13}$$

The equation for computing the optimal value $\omega_j^{\,*}$ is as follows:

$$\omega_j^{\,*} = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{14}$$

3.2.2. Machine learning interpretive models: SHAP and ALE

In this study, we use SHAP and ALE models to explain the XGBoost model. SHAP can interpret the outcomes of each sample (local) and all samples (global), while ALE only takes the average results of all samples (whole). Specifically, the SHAP value and SHAP interaction value of each feature are calculated to represent the effects of single variables and combination variables on PCEE, respectively, and the ALE of each feature is used to reflect the function of the effects of single variables on PCEE.

SHAP values are derived from concepts found in game theory and game theory has theoretically demonstrated that the SHAP value is the ideally distinct solution. A credit distribution procedure can be used to explain a forecast. The prediction is seen as the completion of the project, with each explanatory feature acting as a "worker" within it. The SHAP value represents how equally the successes among the explanatory aspects are credited.

This study used the feature attribution methods to learn more about how individuals assess review helpfulness:

$$g(z^{'}) = \phi_0 + \sum_{i=1}^{9} \phi_i z_i^{'} \tag{15}$$

where g represents the linear explanatory model. ϕ_0 is a scalar coefficient. ϕ_i is the feature attribution value of feature i. $z_i^{'}$ shows whether i is monitored ($z_i^{'} = 1$ if monitored; $z_i^{'} = 0$ not monitored).

According to the SHAP framework, ϕ_i (i.e., SHAP value) is the feature attribution value in the SHAP (Molnar, 2020). The SHAP values attributed ϕ_i to each feature can be determined as follows:

$$\phi_i = \sum_{S\in N\setminus\{i\}} \frac{|S|!(M-|S|-1)!}{|M|!}[f(S\cup\{i\}) - f(S)] \tag{16}$$

where f represents the structure of the XGBoost model. S denotes the set containing non-zero indexes in $z^{'}$. N is the set of all input features, which contains 9 features extracted in Section 3.1.

After taking into consideration each separate feature effect, interaction effects, i.e., additional combined feature effects, can be further obtained as well. The SHAP interaction value $\phi_{i,j}$between features i and j can be calculated using the Shapley interaction index, which is based on the same principles as SHAP values (ϕ_i) (Meng et al., 2021):

$$\phi_{i,j} = \sum_{S\in N\setminus\{ij\}} \frac{|S|!(M-|S|-2)!}{2(M-1)!}\nabla_{i,j}(S), i \neq j \tag{17}$$

ALE is also a post-model interpretation method applied to machine learning models, i.e., XGBoost in this study. ALE is used to estimate the effect of a single feature (i.e., GPDpc, POP, UDD, UPC, UPR, UPI, ACC, RTD, RGD) on the target variable PCEE while keeping all other variables constant.

The ALE plots for a single numerical feature are estimated as follows (Molnar, 2020):

$$\hat{f}_{j,ALE(x)} = \hat{\hat{f}}_{j,ALE(x)} - \frac{1}{n}\sum_{i=1}^{n}\hat{\hat{f}}_{j,ALE\left(x_j^{(i)}\right)} \tag{18}$$

Table 3
Parameter tuning results of XGBoost.

Parameter	Description	Search range	Selected value
colsample_bytree	The proportion of features randomly selected for each tree	[0.1,1.0]	0.9
gamma	Minimum loss reduction required to further split a leaf node	[0.01,0.07]	0.065
learning_rate	Step size for updating model parameters during each iteration	[0.1,0.5]	0.1
max_depth	Maximum depth of each tree in the ensemble	[3, 10]	7
min_child_weight	The minimum sum of sample weights required in a leaf node	[1, 5]	1
n_estimators	Number of trees in the ensemble	[30,100]	35
reg_alpha	L1 regularization term weight	[0.1,1]	0.1
reg_lambda	L2 regularization term weight	[0.1,1]	0.1
subsample	The proportion of samples used for training each tree	[0.5,1.0]	0.8

where $\widehat{f}_{j,ALE(x)}$ represent the ALE value of feature j. $\widehat{\widehat{f}}_{j,ALE(x)}$ shows the uncentered effect of feature j. $\widehat{\widehat{f}}_{j,ALE}\left(x_j^{(i)}\right)$ is the uncentered effect of feature j in sample i.

The ALE function estimates the accumulated effect of the variable on the target variable PCEE over its range. The slope of the ALE function indicates how the expected prediction changes as we move from one interval to the next, which is detailed in Section 4.2. ALE plots (i.e., second-order ALE plots) can show the interaction effect of two features, following the same calculation principles as for a single feature. In-depth explanations of the second-order ALE plot can be found in Section 4.3.

4. Result analysis

The R-squared result of cross-validation of the XGBoost is 0.80, indicating that this model can accurately predict the relationship between the extracted features and per capita carbon emissions. This study is based on a grid search algorithm to select hyper-parameters of the XGBoost model (Putatunda et al., 2018; Tarwidi et al., 2023). Table 3 shows some of the most commonly tuned hyperparameters and the selected value. In the grid search process, all possible combinations of hyperparameters in the search interval are listed to make grids. Then, utilizing the training dataset, the error resulting from each grid is evaluated via the objective function in Eq. (11). The optimal hyperparameters are those with the lowest error.

The results of the XGBoost model will be explained next mainly using the SHAP and ALE models. The results will be analyzed in terms of the importance and heterogeneity effects of features, nonlinear between individual features and PCEE, and interaction effects

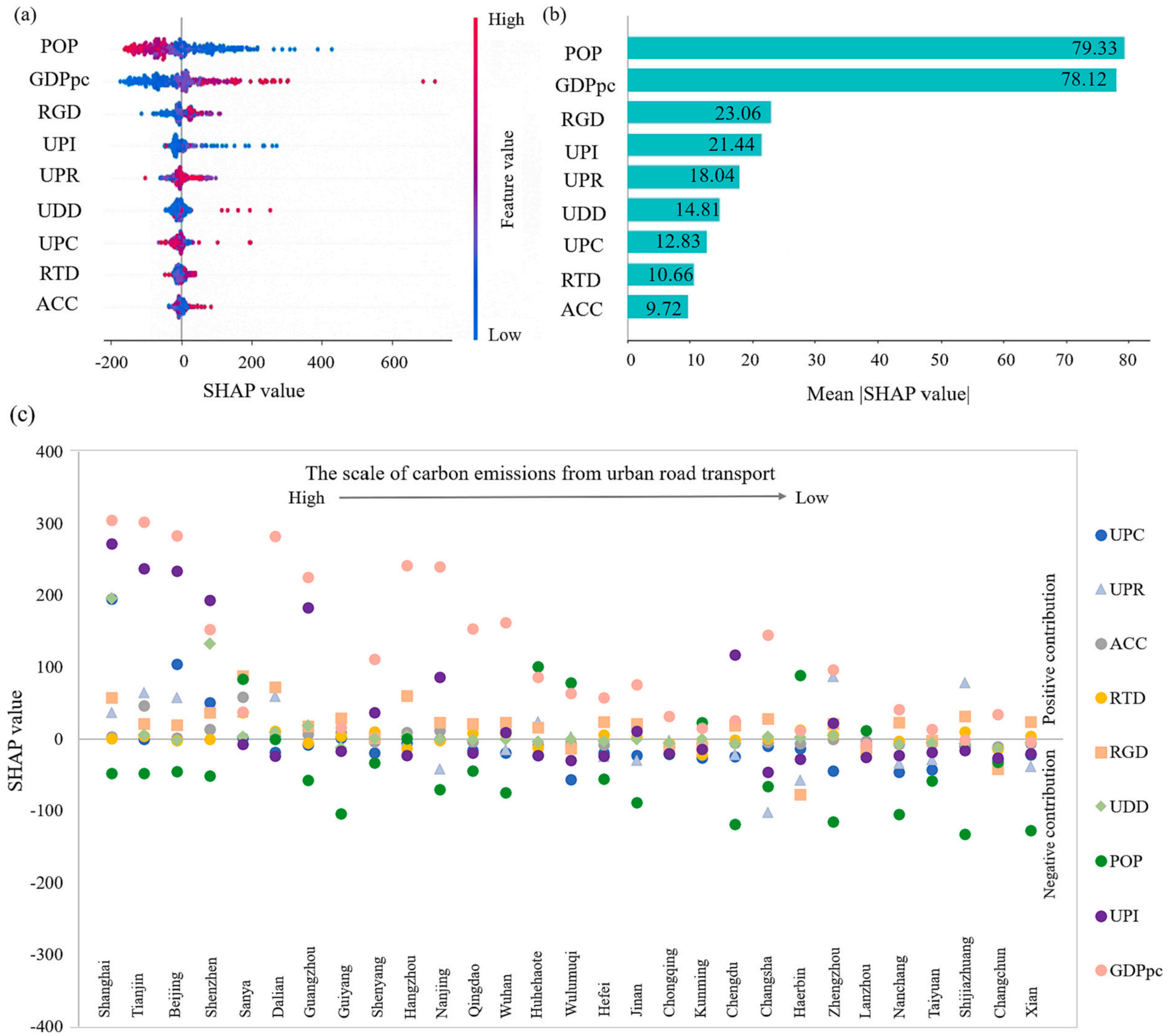

Fig. 2. Contribution of each feature to traffic CO_2 emissions.

of multiple features on PCEE, respectively.

4.1. Drivers of the urban built environment to traffic CO_2 emissions

Fig. 2 shows the relative importance of spatial morphological metrics on per capita traffic emissions PCCE. Fig. 2(a) shows that the population density feature POP is the most important feature, and lower POP has a higher SHAP value, meaning a larger impact. Heihe and Jiuquan (the two blue scatters on the rightmost side of the POP) have low population density, but a higher impact of POP on PCCE. In addition, we statistically find that per capita carbon emissions (i.e., PCCE) in these two cities are higher than in most cities. This is partly due to the small size of the local population and partly due to the high traffic carbon emissions due to the attraction of tourists to the scenic area in Heihe and the launch center in Jiuquan. The importance of GDPpc is different from population density. Higher GDPpc means a larger impact. RGD is an important feature in third place. Fig. 2(b) shows the mean importance of features. POP and GDPpc are the two features that have the greatest impact on traffic CO_2 emissions, and their mean absolute SHAP values are 79.33 and 78.12, respectively. The mean absolute SHAP values of RGD, UPI, UPR, UDD, UPC, RTD, and ACC are 23.06, 21.44, 18.04, 14.81, 12.83, 10.66, and 9.72 respectively.

Fig. 2(c) presents the dominant morphological features across cities. Shanghai has the largest traffic CO_2 emission scale, and Tianjin and Beijing come second. It can be seen that urban morphological features have a heterogeneous impact on per capita traffic emission. In general, the economic level (GDPpc) has the greatest positive contribution to the PCCE, followed by spatial isolation degree (UPI), spatial development degree (UDD), road density (RGD, RTD), and spatial regularity (UPC, UPR). Population density (POP) has a negative contribution to PCCE. These results identified the control priority of urban morphological features, which benefit long-term emission reduction.

The classification of cities according to the main influencing factors facilitates the differentiated management and localization of traffic CO_2 emissions, thus making policies in the classified cities more targeted and leading to more meaningful results. Hence, the percent of the absolute value of each feature' SHAP value in the total value is further calculated for cities, and features with a percentage >50% are identified as the dominant feature of PCCE in that city. As shown in Fig. 3, the cities with POP as the dominant feature are the most, and the least cities with UDD as the dominant feature, such as Qinhuangdao. Cities with POP as the dominant feature are distributed mainly in the northeast and central. Some cities along the southeast coast are dominated by UPC and GDPpc. In cities where POP dominates traffic CO_2 emissions, combined with the results in Fig. 2(a), the higher the value of POP, the lower the value of PCCE, which may be due to the high population density suitable for building high-capacity public transportation. Corresponding policies can be introduced to encourage travelers to use high-capacity public transportation thus reducing CO_2. The value of PCCEs in some cities located in coastal areas is mainly influenced by GDPpc. The role of GDPpc growth in these cities is promoting CO2, which is consistent with the trend of influence of per capita GDP on road energy consumption discovered by Chai et al. (2016).

4.2. The nonlinear relationships between the influencing features and carbon emission

Based on ALE and SHAP models, this section explores the nonlinear relationship between features and PCCE from the overall and individual perspectives, respectively. Specifically, Figs. 4 and 5 show the effect of each feature on the carbon emission PCCE. Fig. 4 shows the ALE plots for each feature, and the one-dimensional ALE analysis indicates the main effect (y-axis) of a feature on the dependent feature (i.e., PCCE) at a certain value, where a plus value represents a positive effect and a minus value represents a negative effect. In contrast to Fig. 4, which shows the overall trend, Fig. 5 shows the impact of each feature on each city from an individual (city) perspective.

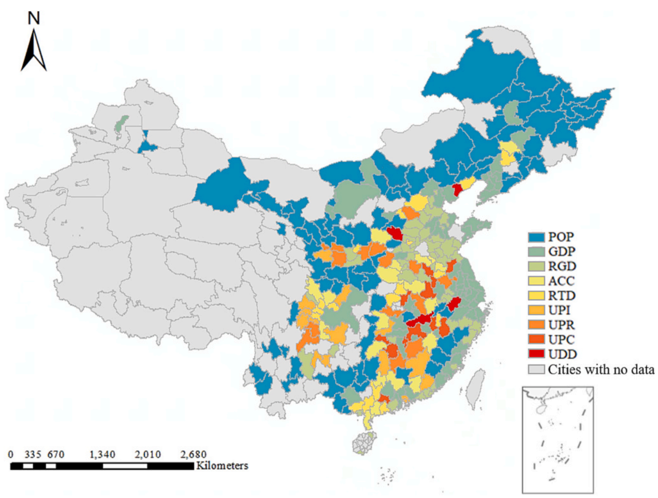

Fig. 3. Classification results of cities according to the main contributor's feature.

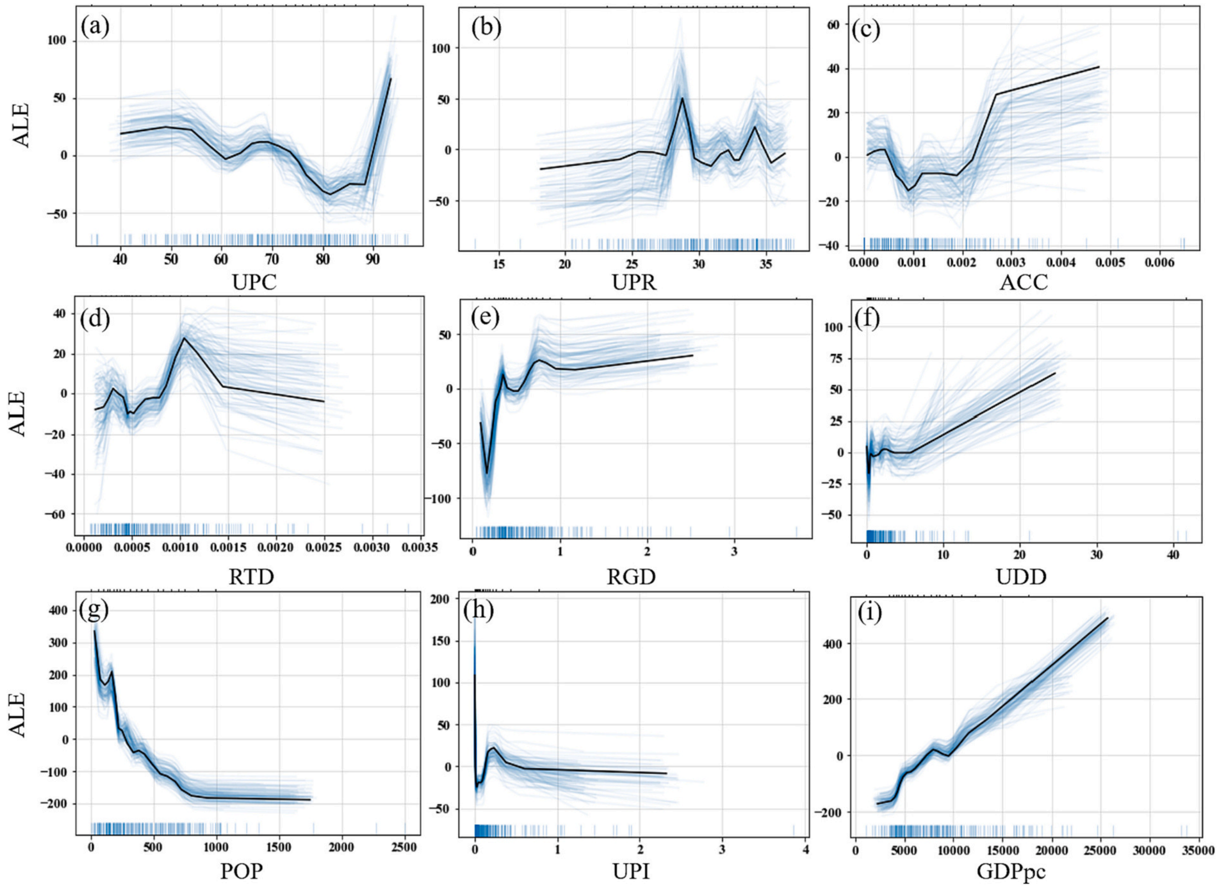

Fig. 4. ALE plot of each feature.

Firstly, the nonlinear relationship between each feature and PCCE can be obtained from Fig. 4, where GDPpc is more like a linear relationship with PCCE. Further, some thresholds can be identified to help develop policies to reduce emissions. In Fig. 4(a), UPC set at 60 is the first dividing line of its positive and negative contributions, and 75 is the second dividing line. In Fig. 4(b), when the value of UPR is in the interval [27, 30], its positive contribution to PCCE is the largest. Similarly, in Fig. 4(c), the dividing line between the positive and negative effects of ACC on PCCE is 0.0022. When ACC is greater than this value, the positive impact on PCCE will gradually increase. When the value of RTD in Fig. 4(d) is in the interval [0.008, 0.0015], the positive effect on PCCE is the largest, and the positive effect decreases after the value of 0.0015. The divider between positive and negative effects in Fig. 4(e) is RGD equal to 0.5. The positive effect of UDD on PCCE increases gradually after its UDD is >7 in Fig. 4(f). A trade-off should be made between Urban development degree and traffic CO_2 emission, and the carbon emission brought by urban development should be considered at the same time. The dividing line between the positive and negative effects of POP on PCCE in Fig. 4(g) is roughly POP equal to 300. In Fig. 4(h), the effect of UPI on PCCE fluctuates when the value is in the interval [0,0.5] and has almost no effect on PCCE when the value is >0.5. In the last subplot, the effect of GDPpc on PCCE increases as it takes on more values, and the dividing line between positive and negative effects is GDPpc equal to 10,000 dollars.

The scatters in Fig. 5 are cities, and the scatter color depends on the size of the PCCE value of that city, the "x" in the figure is the original value taken by the feature, and the "y" is the SHAP value of the feature. Fig. 5(a) shows the UPC's impact on PCCE. When the UPC is between 62% and 78%, it has a significant influence on PCCE. When spatial compactness is >80%, it has a significant effect. However, when the UPC is <60%, its economic level is usually not high, and the impact is not obvious. The greater the UPC between 60% and 75%, the worse the emission reduction potential. Shenzhen, Shanghai, Beijing, and Dongguan's compactness is >94.5%. When is >94.5%, the greater the compactness, the greater the positive impact on PCCE. Fig. 5(e) shows the dependence of PCCE on RGD. When RGD is small (<0.5), it has a negative effect. With the increase of RGD, the negative effect is getting smaller and closer to 0 influence (=0.5). When RGD is larger than 0.5, RGD has a positive impact. The cities with RGD between 0.5 and 1.0 are Huizhou, Sanya, Huzhou, Weihai, Quzhou, Jiangmen, Wuhai, Yantai, Dalian, Hangzhou, and Jinhua. After RGD reaches a certain value (=1.0), the influence is gradually stabilized (SHAP value fluctuates little). The cities with geometrically denser roads tend to have higher POP and have a greater impact on PCCE. Fig. 5(g) shows the nonlinear relationship between the POP and PCCE. The POP has a negative correlation with PCCE in general, and this negative relationship is particularly salient at the value range between 0 and 800 have a sharp downward slope in this value range). When POP is >800, its influence on PCCE gradually stabilizes. The threshold effect

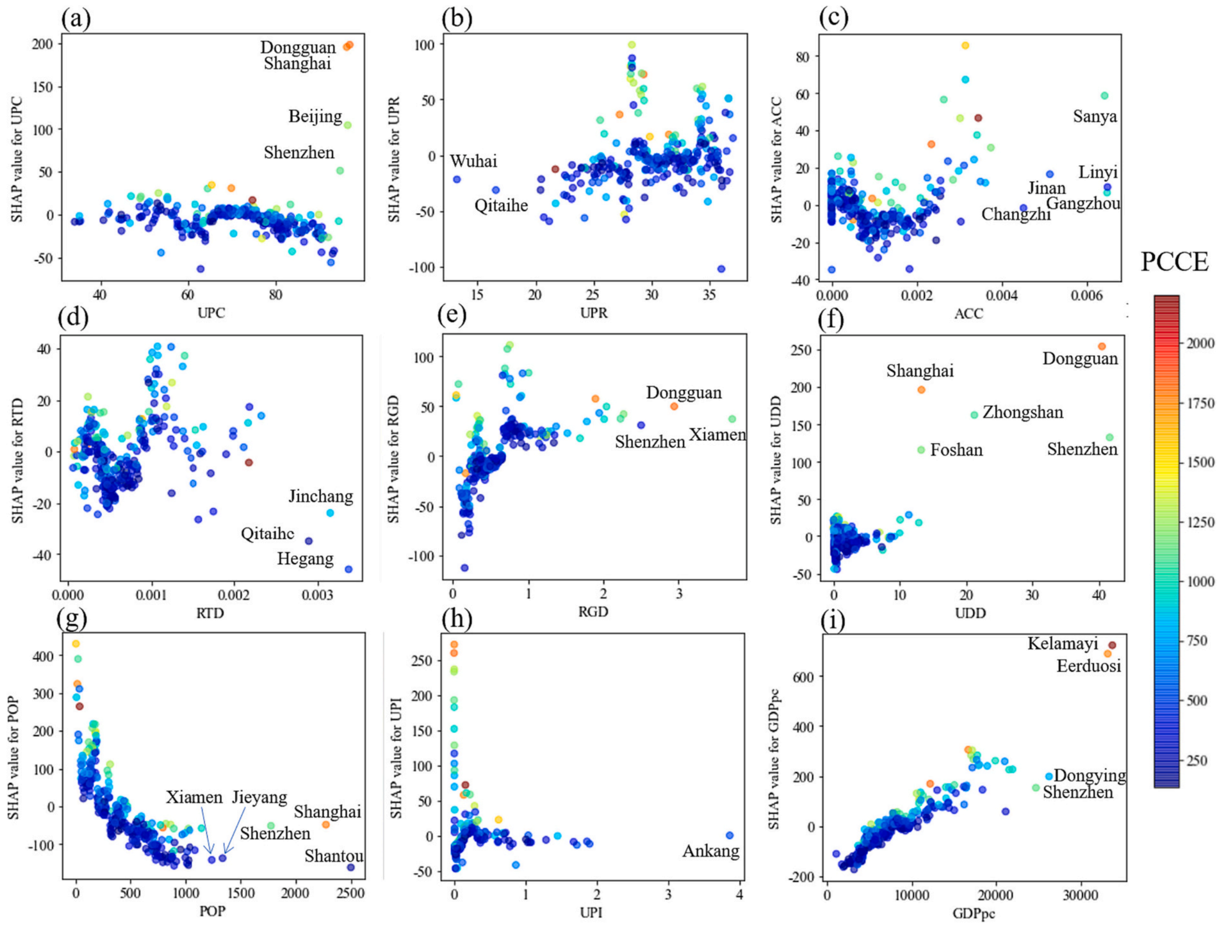

Fig. 5. Scatter plot of the SHAP value of the individual feature on PCCE.

demonstrated in the results indicates that PCCE is more sensitive to changes in population density when the population density is low. Fig. 5(i) shows the influence of different ranges of GDPpc values on PCCE. When the value of GDPpc is >5000 dollars, the higher the GDPpc, the greater the impact on traffic CO_2 emissions; When GDPpc is small, it means that the economy is not good and the travel intensity of residents is low. When the GDPpc is between 5000 and 10,000 dollars, there is a stratification in the scatter, and the influence of the low POP is greater than that of the high POP, which may be caused by the availability of public traffic. Among the cities with GDPpc >10,000 dollars, the cities with high population density are in the majority. Wuxi, Baotou, Guangzhou, Suzhou, Shenzhen, Dongying, Ordos, and Karamay, whose GDPpc is >20,000 dollars. Especially, Ordos and Karamay have the highest GDPpc, but the corresponding Population Density is the lowest, which indicates that GDPpc contributes much more to the PCCE than other cities.

4.3. *Interactive effects between multiple features of the urban built environment*

The analysis in this section primarily relied on two essential techniques: second-order ALE plot and SHAP interaction values. Due to the favorable characteristics of the SHAP method, this section placed particular emphasis on the analysis of results revealed by SHAP interaction values.

Second-order ALE plots are often used to analyze the interaction effect between two features. These plots are based on one-order ALE plots, but they show changes in the model response for different combinations of feature values when considering two features. As shown in Fig. 6, the x-axis represents the value range of GDPpc, and the y-axis represents the value range of the compactness of built-up area UPC. The colors in the legend represent the magnitude of the ALE value, the positive number (red) indicates that the interaction effect of the two features is to promote PCCE, and the negative number (blue) is the opposite. In the ALE plot, there are two sets of percentiles on the x and y axes, representing the distribution of feature values. Combined with the color and percentage (not the area) in the figure, it can be seen that the interaction between GDPpc and UPC is dominated by red, that is, it mainly promotes PCCE. The figure displays two regions where the interaction between GDPpc and UPC suppresses PCCE. One is that the GDPpc is within [10,000,20,000] and the UPC is within [80,90]; the other is that the GDPpc is >7500 and the UPC is <60.

The SHAP interaction values of features pair (i,j) are the higher-order interactions based on the generalization of SHAP values of

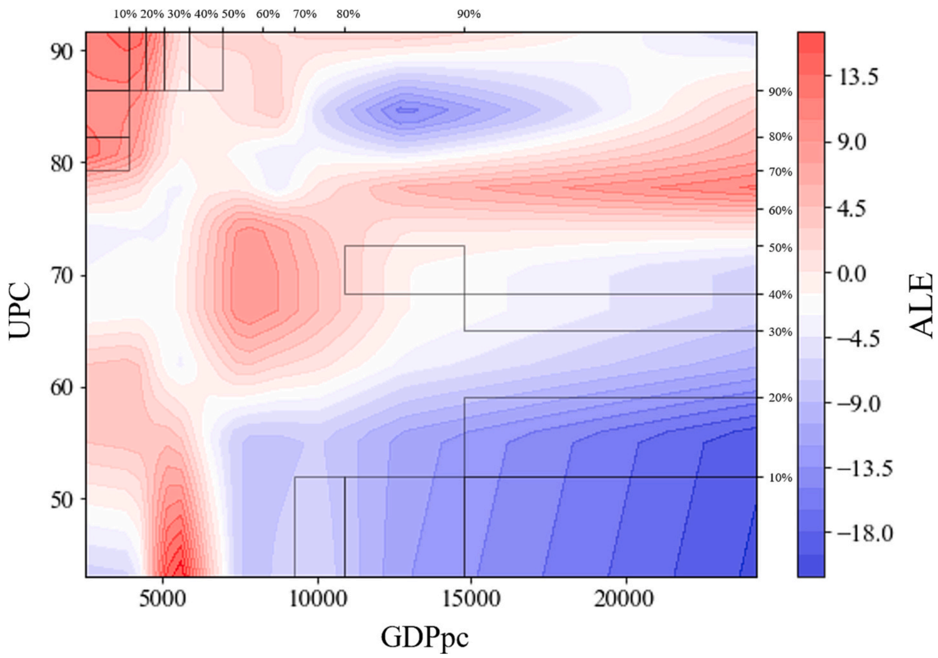

Fig. 6. Second-order ALE plot of features GDPpc and UPC.

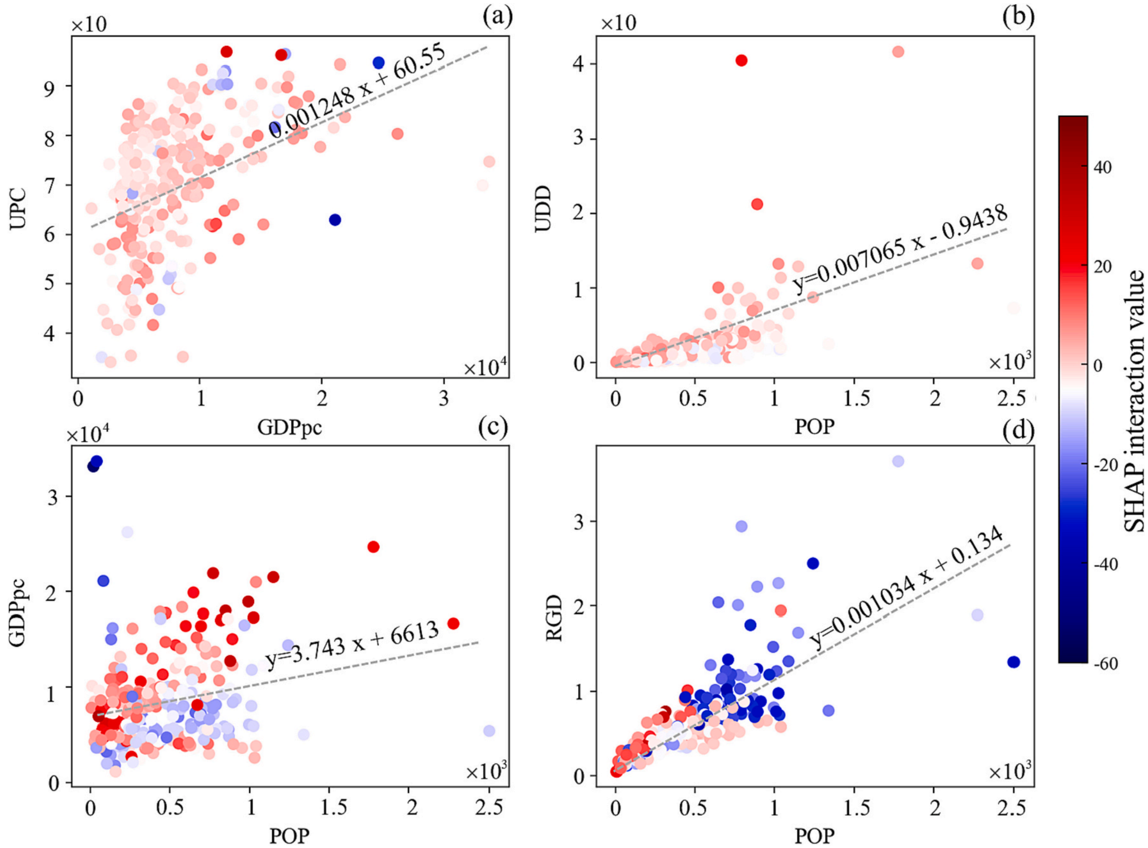

Fig. 7. SHAP interaction value of features.

feature i, which can be obtained by Eq. (17). SHAP can fast exact computation of pairwise interactions of features on each sample, which returns a matrix for every city, where interaction values usually reflect interesting complex effects.

As shown in Fig. 7(a), with the growth of GDPpc, the compactness of built-up area UPC shows an upward trend of change. The color of the scatter is predominantly red, suggesting that the interaction of these two features drives the generation of traffic CO_2 emissions in most cities (SHAP interaction value is positive). Only in a few individual cities has the generation of traffic CO_2 emissions been suppressed, as shown by the blue scatters in the figure, which does not have a uniform arrangement. Similarly, Fig. 7(b) shows the interaction of population density POP and urban development degree UDD on PCCE. First, it can be seen that the values of POP and UDD show a positive relationship. The second is that their SHAP interaction values are also dominated by red points, and it is also the combination of these two features that will promote the production of traffic CO_2 emissions. Fig. 7(c) shows the interaction of population density POP and GDPpc on PCCE. The interaction of these two features has both been beneficial (i.e., positive SHAP interaction value) and detrimental (i.e., negative value) on the CO_2 emissions of different urban transport. Most of the cities with negative impacts (blue scatters) have lower GDP values than most of the cities with positive impacts (red scatters). As shown in Fig. 7(d), the geographical density of urban roads RGD and POP show a similar positive correlation. Similarly, the interaction of these two features has both a positive and negative impact on PCCE. In addition, the cities with positive impact (red scatters) are distributed on the left side of the graph, and the corresponding values of POP are in the range [0,500] and RGD in the range [0,1]. When POP takes a value >500 and RGD takes a value >1, the interaction of these two features will jointly suppress the generation of traffic CO_2 emissions.

To explore the relationship between the interaction of two features and the individual action of the features, this study calculated the difference ΔS between $\phi_{i,j}$ and $\phi_i + \phi_j$ (Eq. (19)). Eight types of cities are determined based on the positive (+) or negative (−) value of their original values and ΔS, which are illustrated in Table 4 detailed. For example, $(+,+) \geq 0$ represents that the sum of these two features acting individually contributes to traffic CO_2 emissions PCCE, and the interaction of the two features also contributes to PCCE and is greater than the sum of the individual effects. In addition, if ΔS takes a value >0 it is said that the two features have a synergistic effect, instead it is antagonistic.

$$\Delta S = \phi_{i,j} - \left(\phi_i + \phi_j\right) \tag{19}$$

As shown in Fig. 8, the antagonistic and synergistic effects of the four groups of features are demonstrated. The red scatters in the figure represent the synergistic effect of this group of features, and the black scatters represent the antagonistic effect. The black and red scatter of these four groups of features in the figure is stratified, indicating that when the values of two features are in different ranges, the antagonism and synergism between the features are different. For example, in Fig. 7(d), when POP ranges from 0 to 300 and RGD ranges from 0 to 0.4, their interaction shows antagonism (black scatter). However, when POP is >300 and RGD is >0.4, their interaction shows synergy (red scatter). In addition, each subfigure in the figure divides all cities into 8 categories, and it is important to focus on the category $(+,+) < 0$, which represents the cities in which two features each contribute to the traffic CO_2, but their interactions attenuate this contributing effect.

5. Discussion

These findings indicate that the configuration of the built-up area and road network have heterogeneous impacts on traffic CO_2 emissions and provide practical insights for appropriate strategies. Firstly, according to the dominant features of traffic CO_2 emissions in each city determined by the absolute value of the SHAP value, managers can determine the main direction of CO_2 reduction in different cities. Specifically, in major provincial capitals and municipalities, GDPpc mainly promotes CO_2 emissions, which may be because cities with high GDPpc have a fast urbanization process and many motor vehicles. Meanwhile, UPI (i.e., urban physical isolation) is the second dominant feature that promotes traffic CO_2 emissions in these first-tier cities, such as Shanghai, Beijing, Shenzhen, and Guangzhou. In these cities, there is a separation between work and residence, where travelers have long commuting distances, thereby contributing to emissions. In summary, for these first-tier cities, where GDPpc and UPI are important influencing factors, public transportation networks should be improved to ensure that there are convenient and efficient public transportation connections between various communities and regions. City administrators should provide a variety of transportation options, including metro, bus, and bike-sharing, to reduce the reliance on private cars and reduce traffic CO_2 emissions. Secondly, based on the nonlinear relationship between the built environment and traffic CO_2 emissions, this study determines the threshold value for each feature, which helps decision makers to determine the reasonable range of urban built environment configuration. Here are the value ranges for several features influencing traffic CO_2 emissions. When UPC is >90%, it promotes traffic CO_2. UPR promotes traffic CO_2 within the range of 27% to 30%. ACC promotes traffic CO_2 when it exceeds 0.0025. RTD promotes traffic CO_2 within the range of 0.0010 to 0.0015. RGD promotes traffic CO_2 when it surpasses 0.5. UDD promotes traffic CO_2 when it goes beyond 7. POP promotes

Table 4
Classification based on SHAP values of features i and j and SHAP interaction values.

Type	$\Delta S \geq 0$		$\Delta S < 0$	
	$\left(\phi_i + \phi_j\right) \geq 0$	$\left(\phi_i + \phi_j\right) < 0$	$\left(\phi_i + \phi_j\right) \geq 0$	$\left(\phi_i + \phi_j\right) < 0$
$\phi_{i,j} \geq 0$	$(+,+) \geq 0$	$(+,-) \geq 0$	$(+,+) < 0$	$(-,+) < 0$
$\phi_{i,j} < 0$	$(+,-) \geq 0$	$(-,-) \geq 0$	$(+,-) < 0$	$(-,-) < 0$

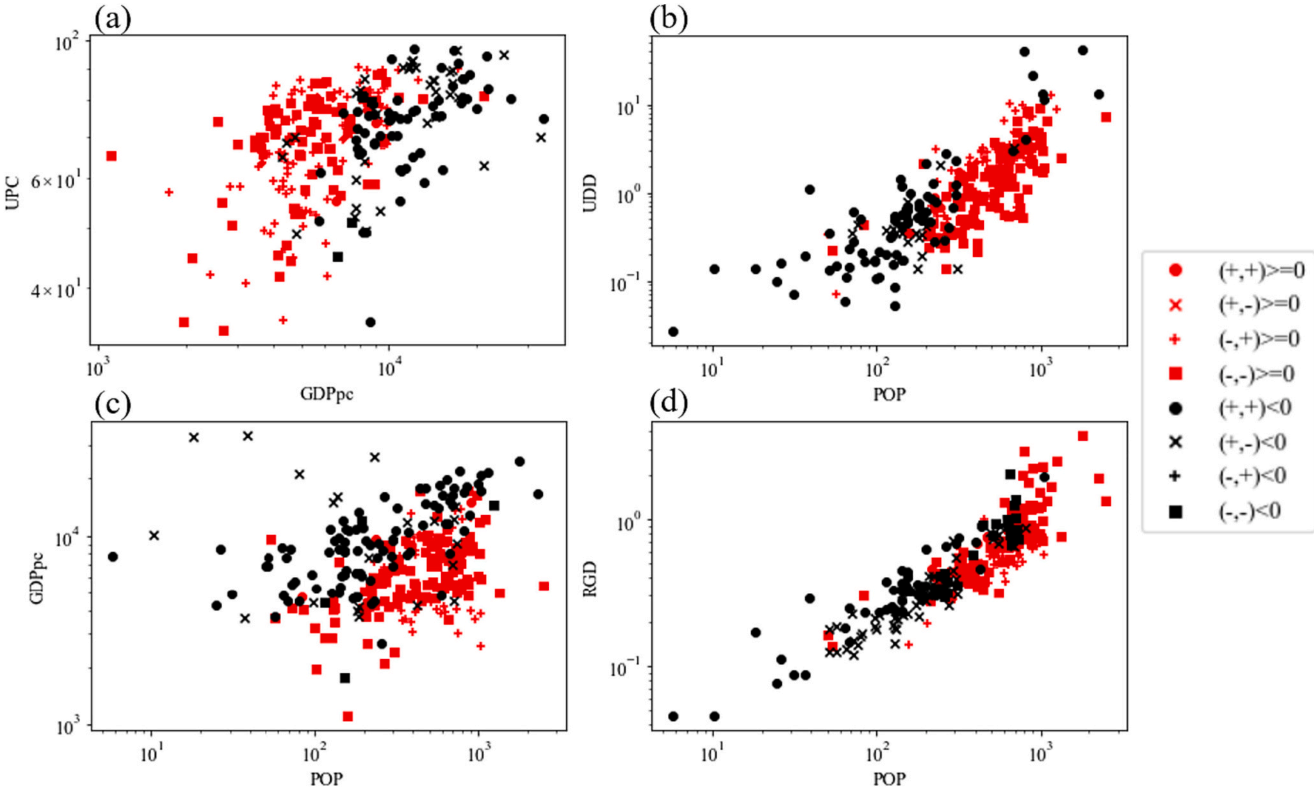

Fig. 8. Antagonistic and synergistic effects of features.

traffic CO_2 when it is higher than 300. These ranges make it possible to check which configuration of features is not justified in each city and to target carbon reduction policies. Finally, due to the antagonistic and synergistic effects of features on transportation CO_2 emissions found based on SHAP interaction value, the effects of multiple features on each other need to be taken into account when developing transportation carbon reduction policies. In cities where GDPpc and UPC exhibit antagonistic effects and both have positive values (i.e., $(+,+) < 0$), if the GDP of such cities will continue to grow, consider adjusting the value of the UPC to weaken the GDP's contribution to traffic CO_2.

Therefore, policymakers should develop city-level mitigation policies informed by the dominant features of different areas, particularly the joint effects of features pair. Urban built-up area and road network configuration interventions will aid in the creation of low-carbon traffic cities.

6. Conclusions

In this study, XGBoost, SHAP, and ALE models of machine learning were used to model the influence of the built-area and road network features on traffic CO_2 emissions (PCEE). The findings show that the same feature has a heterogeneous influence on CO_2 emissions from different urban transport. For major Chinese cities, it is mainly GDPpc that contributes positively to traffic CO_2 emissions and POP that contributes negatively. Overall, the socioeconomic features POP and GDPpc rank first and second in importance, respectively, and the road network feature RGD ranks third, leaving the features of the built-up area with higher importance than the other road network features. Specifically, the mean absolute SHAP values of POP and GDPpc are 79.33 and 78.12, respectively. The mean absolute SHAP values of RGD, UPI, UPR, UDD, UPC, RTD, and ACC are 23.06, 21.44, 18.04, 14.81, 12.83, 10.66, and 9.72 respectively. The results show that socioeconomic features have a larger impact than the built-up area and road network. In addition, the nonlinear relationships between each feature and PCCE were obtained from the overall and individual perspectives, respectively, in this study. The ALE plots from the overall perspective reveal the effects on PCCE under different ranges of feature-taking values. The SHAP value scatter plots from the individual perspective can obtain the impact of each city's feature-taking values on its PCCE. In addition to focusing on the effects of individual features, this study also explored the effects of the interaction of two features on PCCE based on SHAP interaction value. SHAP interaction values for socioeconomic and built environment features are predominantly positive, suggesting that the interaction of these two types of features promotes the production of traffic CO_2. The interaction between POP and GDPpc within socioeconomic features, on the other hand, has positive and negative effects. These two features in the cities with high POP and low GDPpc mainly act as a suppressor of traffic CO_2. The interaction between socioeconomic features POP and road network features RGD is more prominent. The values of their characteristics are also positive and negative, especially in the range of POP$>$300 and RGD $>$ 0.4 mainly suppressing the traffic CO_2.

However, the features affecting traffic CO_2 considered in this study may not be comprehensive enough, and more features could be extracted and applied in future works. With the availability of more information, future research has the potential to study traffic CO_2 from a microscopic perspective by extracting features of the internal composition of the transportation system, such as the number of motor vehicles, the level of public transportation, the proportion of new energy vehicles, and the proportion of shared mobility or truck (Lv et al., 2021; Yang et al., 2022). In addition, future studies can further consider the impact of some policy implementations on emission reduction based on the difference-in-differences model, such as the influence of a new energy car subsidy policy.

CRediT authorship contribution statement

Danyue Zhi: Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Hepeng Zhao:** Investigation, Data curation, Writing - review & editing. **Yan Chen:** Visualization, Writing - review & editing. **Weize Song:** Validation, Funding acquisition, Methodology, Project administration, Writing - review & editing. **Dongdong Song:** Supervision, Data curation, Formal analysis, Investigation, Writing - review & editing. **Yitao Yang:** Formal analysis, Investigation, Supervision, Visualization, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This project is supported by the National Natural Science Foundation of China (No. 20221300672), the Natural Science Foundation of Beijing, China (No. 9204026), and the Tsinghua University Initiative Scientific Research Program (No. 20221080077). The authors thank the anonymous reviewers for their comments and suggestions that helped improve the manuscript.

References

Ashik, F.R., Rahman, M.H., Antipova, A., Zafri, N.M., 2023. Analyzing the impact of the built environment on commuting-related carbon dioxide emissions. Int. J. Sustain. Transp. 17 (3), 258–272.

Axsen, J., Pltz, P., Wolinetz, M., 2020. Crafting strong, integrated policy mixes for deep CO2 mitigation in road transport. Nat. Clim. Chang. 10 (9), 1–10.

Bai, C., Chen, Y., Yi, X., Feng, C., 2019. Decoupling and decomposition analysis of transportation carbon emissions at the provincial level in China: perspective from the 11th and 12th five-year plan periods. Environ. Sci. Pollut. Res. 26 (15), 15039–15056.

Barrington-Leigh, C., Millard-Ball, A., 2017. More connected urban roads reduce US GHG emissions. Environ. Res. Lett. 12 (4).

Boeing, G., 2017. OSMnx: new methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput. Environ. Urban. Syst. 65, 126–139.

Cai, B., Cui, C., Zhang, D., Cao, L., Dai, C., 2019. China city-level greenhouse gas emissions inventory in 2015 and uncertainty analysis. Appl. Energy 253.

Chai, J., Lu, Q.-Y., Wang, S.-Y., Lai, K.K., 2016. Analysis of road transportation energy consumption demand in China. Transp. Res. D-Transp. Environ. 48, 112–124.

Chen, C., 2023. Evaluation methods and optimization strategies for low-carbon-oriented urban road network structures: a case study of Shanghai. Sustainability. 15 (7).

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) 785–794.

Chinese Academy of Sciences, Resources and Environmental Sciences Data Centre, 2020. http://www.resdc.cn/.

City Greenhouse Gas, 2020. China High Resolution Emission Database. www.cityghg.com.

Dikshit, A., Pradhan, B., 2021. Interpretable and explainable AI (XAI) model for spatial drought prediction. Sci. Total Environ. 801.

Duan, Z., Zhao, H., Li, Z., 2023. Non-linear effects of built environment and socio-demographics on activity space. J. Transp. Geogr. 111.

Guan, C., Srinivasan, S., Nielsen, C.P., 2019. Does neighborhood form influence low-carbon transportation in China? Transp. Res. D Transp. Environ. 67, 406–420.

Haase, E., 2022. Driving the environmental extra mile-Car sharing and voluntary carbon dioxide offsetting. Transp. Res. D Transp. Environ. 109.

Han, L., Zhao, J., Gao, Y., Gu, Z., 2022. Prediction and evaluation of spatial distributions of ozone and urban heat island using a machine learning modified land use regression method. Sustain. Cities Soc. 78.

Hoekstra, R., van der Bergh, J., 2003. Comparing structural and index decomposition analysis. Energy Econ. 25 (1), 39–64.

Huang, Y., Zhang, Y., Deng, F., Zhao, D., Wu, R., 2022. Impacts of built-environment on carbon dioxide emissions from traffic: a systematic literature review. Int. J. Environ. Res. Public Health 19 (24).

IEA, 2018. Transport-Tracking Clean Energy Progress.

Kim, Y., Kim, Y., 2022. Explainable heat-related mortality with random forest and SHapley additive exPlanations (SHAP) models. Sustain. Cities Soc. 79.

Kourtit, K., Pele, M.M., Nijkamp, P., Pele, D.T., 2021. Safe cities in the new urban world: a comparative cluster dynamics analysis through machine learning. Sustain. Cities Soc. 66.

Kwak, K.-H., Woo, S.H., Kim, K.H., et al., 2018. On-road air quality associated with traffic composition and street-canyon ventilation: mobile monitoring and CFD modeling. Atmosphere. 9 (3).

Lei, W., Alves, L.G.A., Amaral, L.A.N., 2022. Forecasting the evolution of fast-changing transportation networks using machine learning. Nature. Communications. 13 (1).

Li, F., Cai, B., Ye, Z., et al., 2019. Changing patterns and determinants of transportation carbon emissions in Chinese cities. Energy. 174, 562–575.

Li, Z., Wang, F., Kang, T., et al., 2022. Exploring differentiated impacts of socioeconomic factors and urban forms on city-level CO2 emissions in China: spatial heterogeneity and varying importance levels. Sustain. Cities Soc. 84.

Lin, D., Zhang, L., Chen, C., et al., 2019. Understanding driving patterns of carbon emissions from the transport sector in China: evidence from an analysis of panel models. Clean Techn. Environ. Policy 21 (6), 1307–1322.

Liu, J., Kuang, W., Zhang, Z., et al., 2014. Spatiotemporal characteristics, patterns, and causes of land-use changes in China since the late 1980s. J. Geogr. Sci. 24 (2), 195–210.

Liu, Y., Huang, L., Kaloudis, A., Store-Valen, M., 2017a. Does urbanization lead to less energy use on road transport? Evidence from municipalities in Norway. Transp. Res. D Transp. Environ. 57, 363–377.

Liu, Z., Ma, J., Chai, Y., 2017b. Neighborhood-scale urban form, travel behavior, and CO2 emissions in Beijing: implications for low-carbon urban planning. Urban Geogr. 38 (3), 381–400.

Liu, M., Zhang, X., Zhang, M., Feng, Y., Liu, L., 2021. Influencing factors of carbon emissions in transportation industry based on CD function and LMDI decomposition model: China as an example. Environ. Impact Assess. Rev. 90 (1), 106623.

Liu, Y., Chen, L., Huang, C., 2022. Study on the carbon emission spillover effects of transportation under technological advancements. Sustainability. 14 (17).

Lv, Q., Liu, H., Yang, D., Liu, H., 2019. Effects of urbanization on freight transport carbon emissions in China: common characteristics and regional disparity. J. Clean. Prod. 211, 481–489.

Lv, Y., Zhi, D., Sun, H., Qi, G., 2021. Mobility pattern recognition based prediction for the subway station related bike-sharing trips. Transp. Res. C-Emerg. Technol. 133.

Ma, L., Wu, M., Tian, X., et al., 2019. China's provincial vehicle ownership forecast and analysis of the causes influencing the trend. Sustainability. 11 (14).

Ma, Q., Jia, P., Kuang, H., 2022. Spatial imbalance and factors influencing carbon emission efficiency in China's transport industry. Frontiers. Earth Sci. 10.

Meng, Y., Yang, N., Qian, Z., Zhang, G., 2021. What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values. J. Theor. Appl. Electron. Commer. Res. 16 (3), 466–490.

Mohajeri, N., Gudmundsson, A., French, J.R., 2015. CO2 emissions in relation to street-network configuration and city size. Transp. Res. D Transp. Environ. 35 (Complete), 116–129.

Molnar, C., 2020. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.

Ou, S., He, X., Ji, W., et al., 2020. Machine learning model to project the impact of COVID-19 on US motor gasoline demand. Nat. Energy 5 (9), 666–673.

Park, J., Lee, W.H., Kim, K.T., Park, C.Y., Lee, S., Heo, T.-Y., 2022. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Sci. Total Environ. 832.

Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A., 2020. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accid. Anal. Prev. 136.

Putatunda, S., Rama, K., Acm, 2018. A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost. Paper presented at: International Conference on Signal Processing and Machine Learning (SPML); 2018 Nov 28–30. Shanghai, PR China.

Rong, P., Zhang, L., Qin, Y., Xie, Z., Li, Y., 2018. Spatial differentiation of daily travel carbon emissions in small- and medium-sized cities: An empirical study in Kaifeng, China. J. Clean. Prod. 197, 1365–1373.

Salvia, M., Reckien, D., Pietrapertosa, F., Eckersley, P., Heidrich, O., 2021. Will climate mitigation ambitions lead to carbon neutrality? An analysis of the local-level plans of 327 cities in the EU. Renew. Sust. Energ. Rev. 135, 110253.

Scalfani, V.F., Patel, V.D., Fernandez, A.M., 2022. Visualizing chemical space networks with RDKit and NetworkX. Journal of. Cheminformatics. 14 (1).

Shi, F., Liao, X., Shen, L., Meng, C., Lai, Y., 2022. Exploring the spatiotemporal impacts of urban form on CO2 emissions: evidence and implications from 256 Chinese cities. Environ. Impact Assess. Rev. 96.

Shin, J., Son, S., Cha, Y., 2022. Spatial distribution modeling of customer complaints using machine learning for indoor water leakage management. Sustain. Cities Soc. 87.

Song, W., Zhang, X., An, K., Yang, T., Li, H., Wang, C., 2021. Quantifying the spillover elasticities of urban built environment configurations on the adjacent traffic CO2 emissions in mainland China. Appl. Energy 283.

Sun, D., Zhang, Y., Xue, R., Zhang, Y., 2017. Modeling carbon emissions from urban traffic system using mobile monitoring. Sci. Total Environ. 599, 944–951.

Tarwidi, D., Pudjaprasetya, S.R., Adytia, D., Apri, M., 2023. An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. MethodsX. 10, 102119.
Tu, M., Li, W., Orfila, O., Li, Y., Gruyer, D., 2021. Exploring nonlinear effects of the built environment on ridesplitting: evidence from Chengdu. Transp. Res. D Transp. Environ. 93.
Wang, S., Liu, X., Zhou, C., Hu, J., Ou, J., 2017. Examining the impacts of socioeconomic factors, urban form, and transportation networks on CO2 emissions in China's megacities. Appl. Energy 185, 189–200.
Wang, S., Zeng, J., Huang, Y., Shi, C., Zhan, P., 2018. The effects of urbanization on CO2 emissions in the Pearl River Delta: a comprehensive assessment and panel data analysis. Appl. Energy 228, 1693–1706.
Wang, S., Shi, C., Fang, C., Feng, K., 2019. Examining the spatial variations of determinants of energy-related CO2 emissions in China at the city level using geographically weighted regression model. Appl. Energy 235, 95–105.
Wang, X., Wang, L., Wang, S., Chen, J.-f., Wu, C., 2021. An XGBoost-enhanced fast constructive algorithm for food delivery route planning problem. Comput. Ind. Eng. 152.
Wei, M., Fang, S.J., Chen, S., Lin, R.Y., Huang, Y., Yang, L., 2022. Resilience assessment of road networks in the extremely severe disaster areas of the Wenchuan earthquake. Front. Earth Sci. 10.
Wu, X., Tao, T., Cao, J., Fan, Y., Ramaswami, A., 2019. Examining threshold effects of built environment elements on travel-related carbon-dioxide emissions. Transp. Res. D Transp. Environ. 75, 1–12.
Xu, Q., Yang, R., Dong, Y.-X., Liu, Y.-X., Qiu, L.-R., 2016. The influence of rapid urbanization and land use changes on terrestrial carbon sources/sinks in Guangzhou, China. Ecol. Indic. 70, 304–316.
Xu, Y., Yan, X., Liu, X., Zhao, X., 2021. Identifying key factors associated with ridesplitting adoption rate and modeling their nonlinear relationships. Transp. Res. A Policy Pract. 144, 170–188.
Yang, H., 2003. Research of Simulation Framework and Implement Strategy on Driver, Vehicle, Highway and Environment System (in Chinese). Doctoral thesis.
Yang, W., Li, T., Cao, X., 2015. Examining the impacts of socio-economic factors, urban form and transportation development on CO2 emissions from transportation in China: a panel data analysis of China's provinces. Habitat Int. 49, 212–220.
Yang, L., Wang, Y., Lian, Y., Han, S., 2020. Factors and scenario analysis of transport carbon dioxide emissions in rapidly-developing cities. Transp. Res. D Transp. Environ. 80.
Yang, Y., Jia, B., Yan, X.-Y., Jiang, R., Ji, H., Gao, Z., 2022. Identifying intracity freight trip ends from heavy truck GPS trajectories. Transp. Res. C-Emerg. Technol. 136.
Yu, S., Chen, Z., Yu, B., et al., 2020. Exploring the relationship between 2D/3D landscape pattern and land surface temperature based on explainable eXtreme gradient boosting tree: a case study of Shanghai, China. Sci. Total Environ. 725.
Zhang, M., Mu, H., Ning, Y., Song, Y., 2009. Decomposition of energy-related CO 2 emission over 1991–2006 in China. Ecol. Econ. 68 (7), 2122–2128.
Zhang, R., Matsushima, K., Kobayashi, K., 2018. Can land use planning help mitigate transport-related carbon emissions? A case of Changzhou. Land Use Policy 74, 32–40.
Zhang, L., Chen, H., Li, S., Liu, Y., 2023. How road network transformation may be associated with reduced carbon emissions: An exploratory analysis of 19 major Chinese cities. Sustain. Cities Soc. 95.
Zhou, F., Li, L., 2014. Analysis on the effect of technological innovation diffusion based on the carbon emission reduction. In: Paper Presented at: 13th International Symposium on Management Science and Engineering; 2014. Jan 17–19. Nanjing City, PEOPLES R CHINA.
Zhou, S., Liu, Z., Wang, M., Gan, W., Zhao, Z., Wu, Z., 2022a. Impacts of building configurations on urban stormwater management at a block scale using XGBoost. Sustain. Cities Soc. 87.
Zhou, X., Wang, H., Huang, Z., Bao, Y., Zhou, G., Liu, Y., 2022b. Identifying spatiotemporal characteristics and driving factors for road traffic CO2 emissions. Sci. Total Environ. 834.
Zhu, C., Du, W., 2019. A research on driving factors of carbon emissions of road transportation industry in six Asia-Pacific countries based on the LMDI decomposition method. Energies. 12 (21).