Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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

Inferring freeway traffic volume with spatial interaction enhanced betweenness centrality

Beibei Zhang [a,b], Shifen Cheng [a,b,*], Peixiao Wang [a,b], Feng Lu [a,b,c,d]

[a] State Key Laboratory of Resources and Environmental Information System, IGSNRR, Chinese Academy of Sciences, Beijing 100101, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] The Academy of Digital China, Fuzhou University, Fuzhou, China
[d] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China

ARTICLE INFO

Keywords:
Freeway traffic inference
Spatial interaction
Economic development indicator
Betweenness centrality

ABSTRACT

Freeway traffic volume is strongly correlated with the intensity of regional socioeconomic spatial interactions and the road network structure. Although existing studies have proposed indicators of betweenness centrality (BC) integrated into regional spatial interactions, the socio-economic drivers of freeway traffic volume formation have been neglected. More importantly, existing studies have not established a non-linear response relationship among BC, city socio-economic spatial interactions, and road traffic volume, which severely limits the comprehensive quantification of the role of freeway traffic flow drivers. Therefore, this study proposes a freeway traffic volume inference method that integrates spatial interaction to enhance BC. First, the socioeconomic factors of the origin and destination cities are incorporated into the BC indicator to create an enhanced betweenness centrality indicator (ODBC), which quantifies the strength of spatial interactions between cities. Second, a machine learning approach is used to develop the non-linear response relationship between ODBC and freeway traffic flow to accurately infer traffic volume. Finally, utilizing the SHapley additive explanation approach, the role vectors of intercity freeway traffic volume drivers are quantified. Experiments conducted on data from freeway toll stations demonstrate that the proposed method surpasses the baseline method based on BC and weighted by BC considering only the potential destination or origin city attractiveness, with an improvement in R^2 of 14%, 4.2%, and 4%, and a maximum reduction in RMSE of 40%, 24.5%, and 26%. The proposed method yields higher accuracy for unknown road segments and is easily interpretable.

1. Introduction

Freeways are efficient thoroughfares that link cities and serve as a crucial means of road transportation. The rapid development of freeways has bestowed convenience upon the national economy and regional progress, playing an indispensable role in social and economic growth (Song et al., 2021). Accurate extrapolation of freeway traffic volume helps improve road transport efficiency, which is a crucial concern in current intelligent transport and management (Cheng et al., 2021, 2018; Wang et al., 2023a; Wang et al., 2022a, Wang et al., 2022b).

Existing literature demonstrates that intercity freeways have a substantial impact on enabling the transportation of individuals and goods between cities (Zhao et al., 2024, 2023). The traffic flow on intercity freeways is strongly influenced by the intensity of inter-city interactions and the layout of the road network (Thompson et al., 2019; Wen et al., 2017). Understanding and accurately inferring freeway traffic volume is crucial for optimizing transportation efficiency, improving road safety, and facilitating economic development (Chen et al., 2024). Recent research proposes a freeway traffic volume inference method by considering the appeal of possible destination cities and underscores the significance of the intensity of socio-economic interactions between regions (Zhang et al., 2023). However, this method ignores the impact of road network configuration on the movement of vehicles. Betweenness centrality (BC) has been widely used as a crucial metric in network analysis to assess the significance of edges within a network (Turner, 2007; Li et al., 2020a; Petridis et al., 2020; Pazoky and Pahlavani, 2021; Kazerani and Winter, 2009; Wang et al., 2023b). Road segments with higher betweenness centrality values are inferred to have a greater

[*] Corresponding author at: State Key Laboratory of Resources and Environmental Information System, IGSNRR, Chinese Academy of Sciences, Beijing 100101, China.

E-mail address: chengsf@lreis.ac.cn (S. Cheng).

https://doi.org/10.1016/j.jag.2024.103818
Received 9 January 2024; Received in revised form 25 February 2024; Accepted 2 April 2024

impact on traffic volume, often serving as crucial connectors between different parts of the network. Consequently, betweenness centrality is considered a valuable indicator for comprehending traffic patterns and inferring traffic volume (Henry et al., 2019; Zhang et al., 2022a). However, conventional BC solely emphasizes the network structure (Gao et al., 2013), which hinders a comprehensive representation of the intensity of spatial relationships among regions. Consequently, a recent study proposes a new BC metric that incorporates spatial interactions into the network topology to model the road traffic flow (Wu et al., 2022a). However, this study fails to consider the socio-economic levels of involved cities when assessing the intensity of spatial interactions. This limitation hinders the thorough quantification of the factors driving intercity freeway traffic. More importantly, previous research mostly concentrates on the association between BC or enhanced BC and traffic flow. The nonlinear response relationship among BC, city spatial interaction, and road traffic flow has not been established, thus preventing accurate extrapolation of intercity freeway traffic flow.

This study proposes a freeway traffic volume inference method with an enhanced BC metric (Origin-Destination interaction embedded BC, ODBC) from the formation mechanism perspective, considering the strength of socio-economic spatial interactions among cities and the effect of traffic network configuration on freeway traffic volume. The following are this study's primary contributions:

(1) The socio-economic levels of cities are used to build an enhanced BC, allowing for a quantitative representation of the strength of spatial interactions among cities on road network structure.

(2) A machine learning approach is proposed to model the relationship among network structure, spatial interactions, and freeway traffic, chosen for its capability to capture complex patterns and interactions inherent in the data, thereby enhancing the precision of inferring freeway traffic flow.

(3) The Shapley additive explanation approach (SHAP) is employed to quantify the influence of ODBC on freeway traffic volume, enhancing our understanding of freeway traffic volume formation mechanisms by providing interpretable insights into the contributions of different factors.

2. Methods

The research framework of this research is shown in Fig. 1. Firstly, a city socio-economic interaction enhanced BC indicator by introducing the origin and destination cities attractiveness is developed in Section 2.1. Secondly, a freeway traffic volume inference method based on ODBC using various machine learning methods is constructed in Section 2.2. Finally, SHAP is used to quantify the action vectors of factors and the interactions between various factors on freeway traffic volume.

2.1. Spatial interaction enhanced betweenness centrality

To measure the socio-economic interactions between candidate origin and destination cities on freeway traffic, this study develops an enhanced BC. Firstly, the BC values of road segments are computed. Specifically, the freeway directed weight network $G(V,E)$ is constructed using the road network topology. V represents the collection of nodes, while E denotes the set of road segment edges. The road segment is oriented from its starting node to its ending node, and the weight is the road length. The BC value of road segment e represents the mediating role of road segment e in the freeway network, i.e. the extent to which it

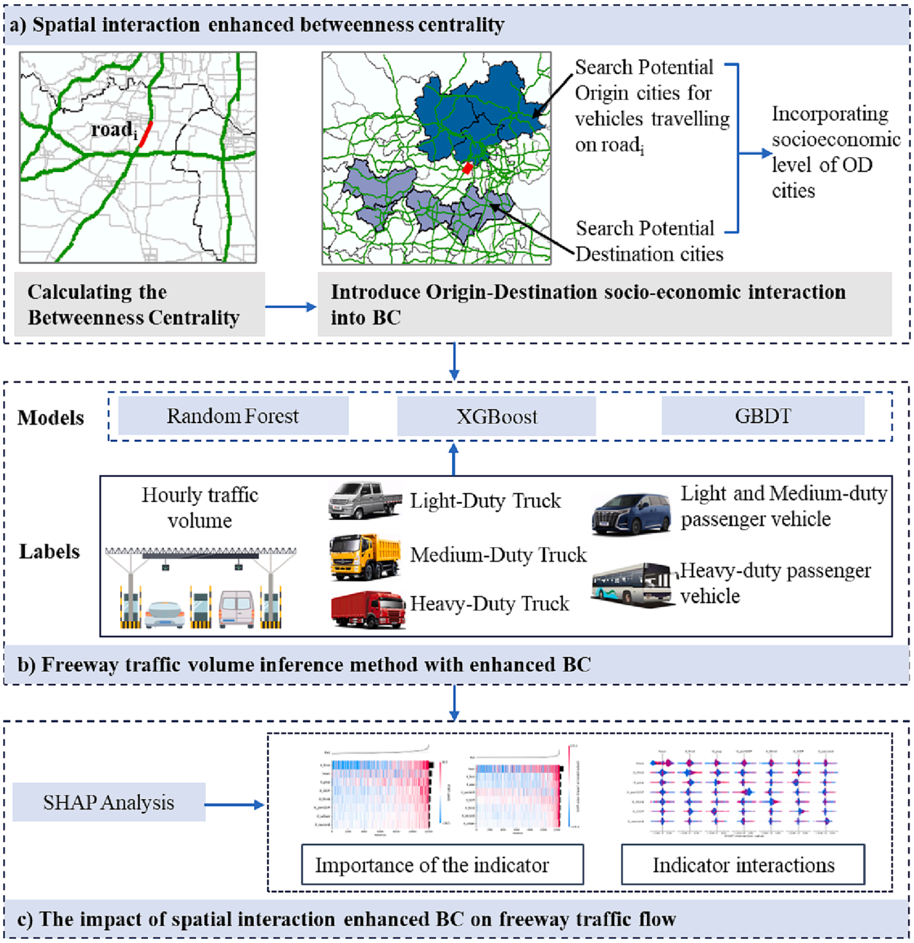
a) Spatial interaction enhanced betweenness centrality
road$_i$
Search Potential Origin cities for vehicles travelling on road$_i$
Search Potential Destination cities
Incorporating socioeconomic level of OD cities
Calculating the Betweenness Centrality
Introduce Origin-Destination socio-economic interaction into BC
Models
Random Forest
XGBoost
GBDT
Labels
Hourly traffic volume
Light-Duty Truck
Medium-Duty Truck
Heavy-Duty Truck
Light and Medium-duty passenger vehicle
Heavy-duty passenger vehicle
b) Freeway traffic volume inference method with enhanced BC
SHAP Analysis
Importance of the indicator
Indicator interactions
c) The impact of spatial interaction enhanced BC on freeway traffic flow

Fig. 1. The research framework. (OD represents Origin-Destination. BC represent Betweenness Centrality.).

controls the traffic exchanges between cities (Yang et al., 2022).

Secondly, previous research has demonstrated a robust correlation between urban socioeconomic indicators and intercity truck volume (Zhang et al., 2023; Li et al., 2020b). Specifically, GDP and GDP per capita offer insights into economic activity and prosperity levels, while population density provides an indication of urban density and potential demand for goods and services. Additionally, the ratios of primary, secondary, and tertiary industries shed light on the economic structure and diversification of urban areas, which can impact freight movements. Finally, the urbanization rate reflects the degree of urban development and its associated infrastructure demands. By incorporating these indicators, we aim to capture diverse aspects of urban socioeconomic dynamics that may influence intercity truck volume. Therefore, seven easily accessible urban socioeconomic indicators of candidate origin and destination cities have been selected to further enhance BC indicators. The calculation as shown in Equation (1).

$$ODBC_{e,m} = \sum_{s,t\in V; s\neq t} \sum_{i,j\in C; i\neq j} \frac{\sigma_{s,t}(e)}{\sigma_{s,t}} \frac{O_{e,i,m} D_{e,j,m}}{d_{OD,e,ij}} \tag{1}$$

where $ODBC_{e,m}$ represents enhanced BC value by city spatial interactions of road e; $\sigma_{s,t}(e)$ represents the shortest paths between s and t; $\sigma_{s,t}$ denotes the total count of shortest pathways between s and t in this network. $O_{e,i,m}$ and $D_{e,j,m}$ represent the m statistical indicators of the potential origin $O_{e,i}$ and the potential destination city $D_{e,j}$ passing through road segment e; m represents the statistical indicators of GDP, population, industrial structure, and the rate of the urbanization, respectively.

The specific steps for identifying the potential origin city $O_{e,i}$ and potential destination city $D_{e,j}$ of the road segment e are (1) setting the road segment e as the starting location in the national freeway directed topology network $G(V,E)$ to identify the potential destination city $D_{e,j}$ of the vehicle with e as the starting position as well as the shortest connectivity distance $d_{D,j}$ from the road segment e to the $D_{e,j}$; (2) setting the road segment e as the termination location in the national freeway directed topology network $G(V,E)$ to identify the potential departure city $O_{e,i}$ as well as the shortest connectivity distance $d_{O,i}$. Both the potential origin city $O_{e,i}$ and the destination city $D_{e,j}$ are searched for in the set of cities C in mainland China. $d_{OD,ij}$ denotes the shortest distance between the potential origin $O_{e,i}$ and the potential destination city $D_{e,j}$ passing through road segment e.

2.2. Freeway traffic volume inference with enhanced BC

The formation of freeway traffic is inextricably linked to the socio-economic interaction of cities (Yang et al., 2023). Similar to recent research (Zhang et al., 2023), this study uses seven easily accessible socio-economic indicators as independent variables (Table 1). These socio-economic indicators are weighted using Equation (2) in Section 2.1 to calculate the ODBC indicator of the road on which each toll station is situated as the independent variables. Considering the time-dependence of traffic flow, hourly time indicator is introduced.

Table 1
Influence factors of freeway traffic volume.

Category	Independent variables	Variable code
Time	Hour	hour
Socio-economic interactions in origin and destination cities	GDP	GDP
	Population density	popdensity
	Primary industry ratio	first
	Secondary industry ratio	second
	Tertiary industry ratio	third
	GDP per capita	perGDP
	Urbanization rate	urban

The dependent variables comprised vehicle type-categorized hourly traffic flow statistics. The passenger vehicles include Light and Medium-duty passenger vehicle (LMPV) and Heavy-duty passenger vehicle (HPV). Trucks include Light-duty truck (LDT), Medium-duty truck (MDT) and Heavy-duty truck (HDT). The basis of vehicle classification (Wu et al., 2022b) is shown in Table S1. The detailed description of the data is in Section 3.1.1.

This study utilizes the Random Forest method (RF) to develop the relationship between road segment ODBC and traffic volume for each vehicle type. Random forest is a classical machine learning technique that incorporates an integrated learning idea and a decision tree classifier, which has the characteristics of high training speed, low possibility of overfitting, and easy operating process (Breiman, 2001). Two widely used machine learning methods, i.e., GBDT (Gradient Boosting Decision Tree) and XGBoost (Extreme Gradient Boosting Tree) methods are also employed for comparison. GBDT is an iterative decision tree algorithm that belongs to the Boosting algorithm, where the final prediction output is obtained by boosting a weak learner to a strong one (Friedman, 2002; Jerome H. Friedman, 2001). XGBoost is an improvement of GBDT that achieves faster convergence and mitigates the risk of overfitting (Yi et al., 2021).

2.3. Analysis of influencing factors on traffic volume

SHAP is used to quantify the effect vectors of each influencing factor on the traffic volume inference model. SHAP calculates the magnitude of the influence that each feature of the sample exerts on the dependent variable and yields the Shapely value (Lundberg et al., 2018; Lundberg and Lee, 2017). In addition, this study measures the interaction of the influencing factors using SHAP, which is used to quantify whether the combined effect of the influencing factors enhances or weakens the impact on traffic volume. Equation (2) represents the calculation of shapely values. The calculation of factor interaction is shown in Equation (3) (Lundberg et al., 2020).

$$f(x) = \phi_0(f,x) + \sum\nolimits_{i=1}^{m} \phi_i(f,x) \tag{2}$$

where $f(x)$ indicates the predicted value; $\phi_0(f,x)$ indicates the mean predicted value for the dataset; m denotes the number of features; $\phi_i(f,x)$ is the SHAP value for the i-th feature.

$$\Phi_{i,j}(f,x) = \sum_{S\subseteq \mathscr{M}\setminus\{i,j\}} \frac{|S|!(M-|S|-2)!}{2(M-1)!} \nabla_{ij}(f,x,S) \tag{3}$$

where $\mathscr{M}$ represents the count of features; the SHAP interactions of feature i and feature j are equally distributed by each feature, namely, $\Phi_{i,j}(f,x) = \Phi_{j,i}(f,x)$, and the total interaction is equal to $\Phi_{i,j}(f,x) + \Phi_{j,i}(f,x)$.

3. Evaluation

3.1. Experimental design

3.1.1. Data sources

Experimental analyses are conducted using data from national freeway network and socio-economic indicators of Chinese cities, as well as sample data gathered from freeway toll stations in the Beijing-Tianjin-Hebei (BTH), China. The first two parts of the data are used to develop ODBC indicators. The freeway toll station data is used to extract the hourly traffic volume of each vehicle classification.

Specifically, the national freeway network data includes attribute information, such as starting and end locations, road length, and topology. Socio-economic data of Chinese cities are derived from the 2022 China City Statistical Yearbook.

Furthermore, the dataset capturing the vehicles that traversed the freeway toll stations in the BTH region on March 15, 2022, from 0:00 to 24:00 h has been acquired. The total amount of data is 14,318,743 records. The data format is shown in Table 2. The dataset records the name, latitude, and longitude of each toll station, as well as the ID, time, and attribute information of each vehicle passing through the toll stations. After map matching and data preprocessing, the BTH toll stations are matched to the freeway network (Fig. 2). Then, based on the vehicle type classification, this study counts the hourly traffic volume of each vehicle classification at 2080 toll stations, the results as shown in Table 3.

To construct the model, the dataset is split into a training set comprising 70 % of the data for model development, and a test set consisting of the remaining 30 % for validation purposes. To prevent overfitting, 10-fold cross-validation is then employed.

3.1.2. Evaluation metrics

Four frequently used evaluation criteria are utilized to quantify the performance of models, as shown in Eqs. (4), (5), (6), and (7). R^2 is a metric used to quantify the level of model fit. RMSE is utilized to quantify the exact size of the deviation between predicted and true data. MAE is a metric that quantifies the gap between predicted and true values, which can indicate the average size of the prediction errors. MAPE is a metric utilized to assess the accuracy of predictions that expresses the degree of discrepancy between predicted and true values as a percentage. A lower MAPE value indicates a higher level of precision in the model's predictions.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \quad (6)$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \quad (7)$$

3.1.3. Parameter selection

Taking the random forest model as an example, the parameters that need to be tuned include the number of trees (n_estimators), the maximum depth (max_depth), and the maximum number of features (max_features). The process of tuning parameters in a random forest involves specifying an estimated range of values for each parameter, individually optimizing these values, and ultimately selecting the combination that yields the optimal parameter values. The permissible values for the parameter n_estimators vary from 1 to 1000. Maximum depth values are from 1 to 1000. The max_features is chosen either as the square root of the total number of features or as the logarithm of that number. Taking LDT as an example, the trend of RMSE with the values of n_estimators and max_depth is demonstrated as shown in Fig. 3. When n_estimators = 867, max_depth = 2325, the max_features is the square root of the feature number, RMSE is the smallest and can be used as the optimal parameter combination for the model.

Table 2
Recording data from freeway toll stations.

Toll gate identification	Vehicle identification	Vehicle type	Time
S005112001010920020	CF2228_1	HDT	2022–03-15 T00:20:57
G450111002001920010	MKV000_0	LDT	2022–03-15 T14:43:38
S003211001000310010	BW1880_0	LMPV	2022–03-15 T21:51:49
G000411001000410010	F1Z466_0	LMPV	2022–03-15 T23:53:07
…	…		…

3.2. Influence of socio-economic interactions

The impact vectors on freeway traffic volume are quantified and SHAP heatmap is drawn according to the predicted value, arranged in ascending order, as depicted in Fig. 4. The row where f(x) is located indicates the predicted traffic volume for each model. Horizontal coordinates represent instances. Vertical coordinates represent features. The rightmost bar represents the SHAP value.

Hourly time is the most important influencing factor for passenger vehicles and trucks. The SHAP values of the hour variables have the most dispersed distribution of the predicted values of passenger vehicle and truck traffic flow. This indicates that the role of hour variables on traffic flow is more complex, which is related to the fact that freeway traffic volume has strong temporal variability characteristics (Zhang et al., 2022b).

Regarding socio-economic factors, the primary industry ratio of the origin and destination cities holds paramount importance for passenger vehicles. Overall, the SHAP of the primary industry ratio increases as the predicted value increases. Primary industry ratio mainly demonstrates to promote passenger vehicle volume. Similarly, population and GDP of cities show an increase in passenger vehicle volume. In contrast, GDP per capita mainly showed a tendency to decrease passenger vehicle volume. The SHAP values of the tertiary ratio show an increasing trend as the predicted values of traffic flow for LMPV increase, while for HPV, the tertiary ratio shows an increasing traffic volume on all samples.

In terms of LDT, the most important socioeconomic factor is the GDP of cities. GDP has the impact of boosting the flow of LDT traffic. In other words, the greater the GDP of the OD cities, the greater the demand for LDT. In terms of MDT, the primary industry ratio is the most important socio-economic factors. The role of the primary industry ratio on MDT traffic is complex, as high values of the primary industry ratio may be found in regions with low or high traffic volume predictions. The SHAP values of GDP per capita are divided into three groups. In terms of HDT, the primary industry ratio is also the most important socio-economic factor. The dispersed distribution of high SHAP values for the primary industry ratio indicates that the role of the primary industry ratio is complex. While the high values of the secondary industry ratio in the cities of origin and destination mainly show the effect of increasing the traffic flow of HDT.

Further, the SHAP value is decomposed into the main effects and interaction effects of each influence factor. Taking HDT as an example, the results are shown in Fig. 5. The horizontal coordinate denotes the SHAP value. Each point indicates an instance, while the colors indicate the feature values in the vertical coordinate. The feature on the diagonal indicates the main effect of the feature. While non-diagonal is the interaction of one feature with another. Figures in the non-diagonal position with the diagonal in the symmetrical position have the same shape but opposite colors.

Specifically, the main impact of the hourly time shows an increase in HDT traffic volume, which is related to the time-of-day restrictions. For example, HDTs are prohibited from being driven on roads located within the Fifth Ring Road in Beijing from 6:00 to 23:00. Therefore, as the value of the hourly time increases, the traffic flow of HDT shows an increasing trend. In terms of the socio-economic factors, the main effect of the primary industry ratio is a reduction in HDT traffic volume. The interaction of the primary industry ratio with the other factors diminishes the propensity of high primary industry ratio values to lower HDT traffic volume to varying degrees. The high GDP and GDP per capita are mostly reflected in the trend towards increased traffic in HDT, which is weakened by interaction with other socio-economic factors, respectively. A similar role is seen for the urbanization rate. While the high values of the secondary industry ratio show a tendency to increase the HDT traffic volume.

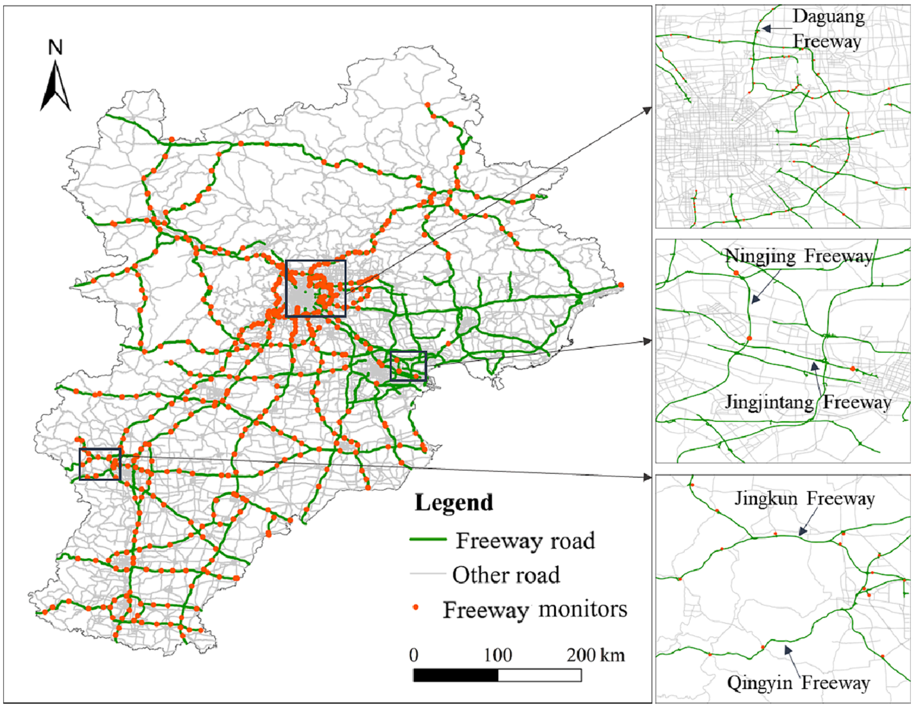

Fig. 2. Distribution of freeways and freeway toll stations in Beijing-Tianjin-Hebei region, China.

Table 3
Hourly traffic volume by vehicle type.

Time	Hour	Road identification	LMPV	HPV	LDT	MDT	HDT
2022–03-15 00:00–00:59	0	5,453,640,000,005	7	0	13	44	137
2022–03-15 12:00–12:59	12	5,454,740,001,857	38	0	22	49	118
2022–03-15 13:00–13:59	13	5,454,650,000,284	41	8	28	8	22
2022–03-15 20:00–20:59	20	5,454,650,000,291	13	9	31	3	14
….	…	…		…			
2022–03-15 23:00–23:59	23	6,255,230,000,044	10	0	0	2	11

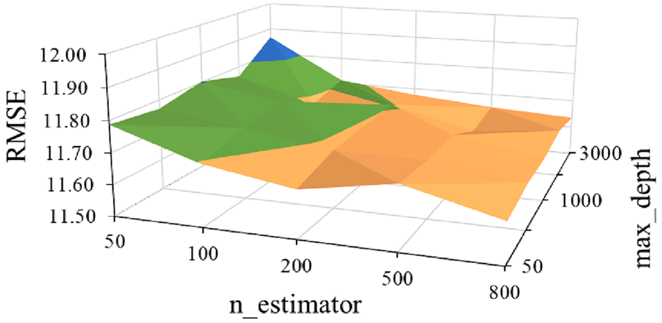

Fig. 3. The trend of RMSE with the values of n_estimators and max_depth.

3.3. Performance results

3.3.1. Inference accuracy

To evaluate the efficacy of the proposed traffic volume inference model, three baseline models are developed. They are baseline-BC based on the BC metric. Baseline-D is a benchmarking model based on BC weighted by the potential destination cities attractiveness. And Baseline-O, a benchmarking model based on BC weighted by the attractiveness of potential departure cities. Specifically, the baseline-BC independent variables include hours and BC. To keep consistency, the proposed model based on the ODBC uses the same socio-economic factors as the Baseline-D and the Baseline-O, with differences in the approach of

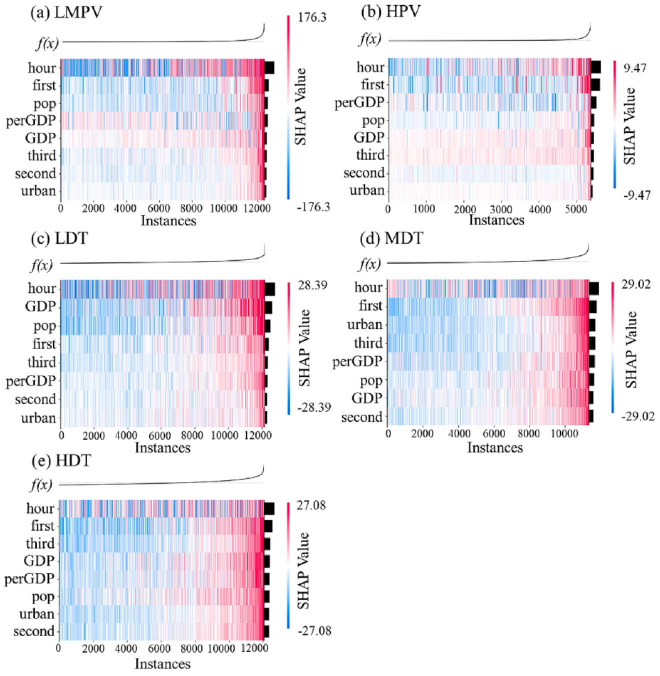

Fig. 4. Heatmap of the role of influence factors on freeway traffic volume.

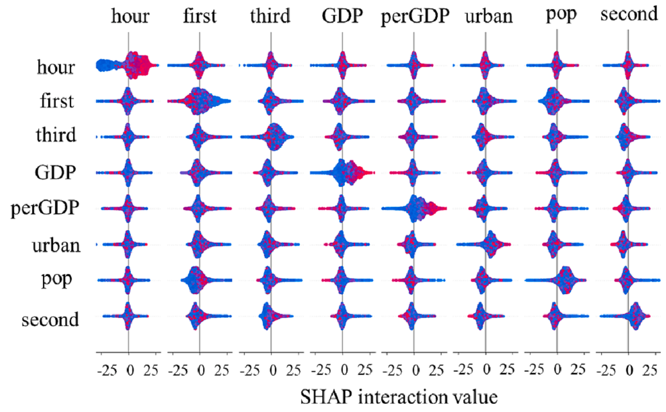

Fig. 5. The interaction effects of each influence factor on HDT traffic volume.

calculating the socio-economic indicator values. Among them, the potential destination city search for Baseline-D aligns with recent literature (Zhang et al., 2023). To better compare the effect of city attractiveness on freeway traffic, only BC values weighted by hour and city socio-economic factors are used. The socio-economic indicators in Baseline-O are weighted using the socio-economic factors of potential departure cities. The potential origin search is in the opposite direction of the potential destination city search in Baseline-D.

The Random Forest, GBDT and XGBoost machine learning methods are used for modelling respectively. Overall, the outcomes of the proposed method surpass those of three benchmark models (Table 4). Comparing the proposed model and the three baseline models, Baseline-O and Baseline-D outperform the Baseline-BC, except for the XGBoost-trained LMPV model effect. This finding suggests that the introduction of city attractiveness helps to increases the precision of freeway traffic volume inference compared to the approach based on the BC of network topology. However, neither Baseline-BC nor Base-O and Base-D have been able to fully capture complex transportation needs. Thus, these benchmark models have some limitations, and the effects are all inferior to the proposed model in this study. Statistically, the proposed model improves R^2 by 0.2 %–23.3 % and reduces RMSE by a maximum of 24 % compared to Baseline-D.

In terms of the evaluation metrics, the R^2 of all models except the HPV is high, with values above 0.8. As far as the MAPE, the proposed model is all lower than the baseline models. However, the absolute value of MAPE in the proposed model is still higher. The reason for this is the existence of some toll stations with small values of hourly traffic volume, which results in the discrepancy between the predicted and actual values being several times the true value, inflating the MAPE value. The LMPV with an hourly traffic average of 137 is used as an example for further validation, and the model is retrained after deleting data with an hourly traffic of 10 or less. The retrained results are displayed in Table S2. Table S2 shows that the MAPE plummets from 0.817 to 0.35. Thus, the high absolute value of the MAPE of the proposed model is strongly linked to the data distribution. In conclusion, the proposed method has high reliability for freeway traffic flow inference.

In addition, RF and GBDT modeling is slightly better than XGBoost and GBDT in comparison to different machine learning methods. Further, RF is faster in training and is easy to operate. In summary, RF model is better than GBDT and XGBoost. Therefore, follow-up modeling is done using RF to compare the results of the baseline and the proposed model.

Table 4
Comparison of prediction accuracy outcomes (in R^2/RMSE/MAE/MAPE) between the proposed method and baselines.

		LMPV	HPV	LDT	MDT	HDT
RF	Baseline-BC	0.897/87/35.8/1.5	0.526/9/2.8/1.1	0.885/17/11/0.9	0.899/17/9.6/1.0	0.888/18/11.9/1.0
	Baseline-O	0.933/72/29.6/1.2	0.589/9/3.0/1.0	0.936/13/8.5/0.6	0.928/15/8.6/0.9	0.922/16/9.9/0.8
	Baseline-D	0.939/69/28.9/1.2	0.536/9/3.1/1.0	0.933/13/8.7/0.7	0.940/13/8.1/0.7	0.917/16/10.0/0.8
	Our model	**0.943/64/25.7/0.8**	**0.588/8/2.8/1.0**	**0.946/12/7.8/0.5**	**0.946/13/7.7/0.6**	**0.933/14/9.2/0.5**
GBDT	Baseline-BC	0.865/102/45.5/2.1	0.29/11/3.1/1.0	0.848/20/11.3/0.8	0.834/20/12.0/0.9	0.843/22/12.8/0.8
	Baseline-O	0.881/96/32.5/1.1	0.488/9/3.3/1.0	0.903/15/9.5/0.6	0.915/16/10.4/1.2	0.913/17/10.7/0.8
	Baseline-D	0.885/94/33.7/1.7	0.498/9/3.2/1.0	0.924/14/9.1/0.7	0.918/16/9.2/0.6	0.921/16/10.1/0.8
	Our model	**0.934/71/27.6/1.2**	**0.614/8/2.5/0.8**	**0.945/12/7.7/0.5**	**0.92/15/9.2/0.7**	**0.931/14/9.3/0.5**
XGBoost	Baseline-BC	0.926/75/31.2/1.3	0.447/10/3.5/1.2	0.929/13/9.0/0.7	0.913/16/10/1.1	0.904/17/11.6/0.8
	Baseline-O	0.926/76/29.4/1.7	0.492/8/3.0/1.2	0.935/13/8.6/0.7	0.924/15/9.4/1.0	0.914/16/10.8/0.9
	Baseline-D	0.917/80/30.3/1.7	0.473/8/3.0/1.2	0.933/13/8.5/0.7	0.934/14/8.2/0.6	0.910/17/11.0/0.9
	Our model	**0.946/65/27.5/1.2**	**0.52/9/3.4/1.2**	**0.943/12/8.1/0.5**	**0.94/13/8.1/0.6**	**0.928/15/9.6/0.6**

3.3.2. *Stability analysis*

Additionally, the stability of both the benchmark and the proposed models are compared. Taking LMPV as an example. In this scenario, a set of 20 identical random numbers is employed to model the proposed and baseline models, using the RF. Furthermore, all models undergo parameter tuning. The 20 modelling results of the proposed model and the three baseline models are shown in Fig. 6. From Fig. 6, the mean value of R^2 of the proposed method is higher than these benchmark models. Importantly, the distribution intervals and interquartile

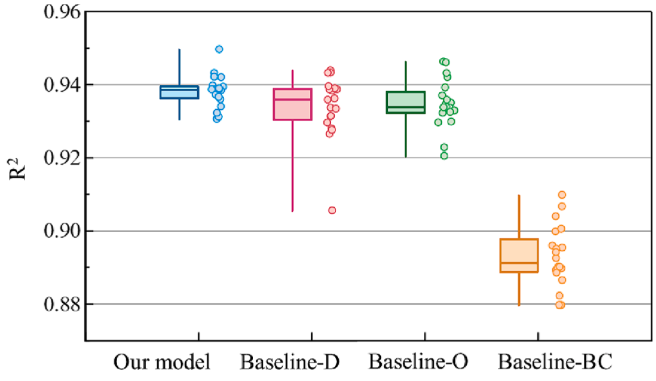

Fig. 6. The stability analysis of the proposed and baseline models.

distances of the proposed model results are smaller than those of the three benchmark models. Specifically, the interquartile range of the proposed model is 0.32–0.47 times that of the baseline models, indicating that the proposed model is more focused and has higher stability.

3.3.3. *Comparative analysis of temporal trends*

For comparing the prediction outcomes on unknown roads, three randomly selected roads have their hourly traffic flow distributions forecasted from 0 to 23 h using both the proposed model and the baseline models. Taking LMPV as an example, the predicted results of different models and the true value distribution are shown in Fig. 7. The prediction outcomes from the proposed model demonstrate closer proximity to the true values and exhibit higher stability (Fig. 7). Statistically, the proposed model surpasses the benchmark models in performance across 61.1 % to 98.6 % of the observed time points. Overall, it proves superior in forecasting the hourly traffic volume of unknown roads compared to the benchmark models.

4. Discussion

Unlike previous research focusing on the correlation between BC and traffic volume, as well as incorporating spatial interaction-BC that did not adequately consider the socio-economic development level of cities, this study expands the application of BC in road traffic flow inference. The results show that freeway inference method with city socio-economic interactions enhanced BC improves R^2 by an average of 14.11 % and reduces RMSE by a maximum of 40 % compared to the baseline method that uses only BC. This finding indicates that city socio-economic interaction enhanced BC helps to improve road traffic inference results. This assertion is supported by the existing research (Li et al., 2021), which suggests that urban industrial development indicators have the capacity to forecast 39.08 % of intercity mobility, whereas variables such as GDP, population density, and urbanization rate accounted for 16.99 % of the prediction. What’s more, the ODBC metrics proposed in this study, whose data are publicly and easily accessible, have some generalization ability and can be applied to traffic flow inference studies in other study areas.

The new proposed method with city socio-economic interaction enhance BC in this study effectively improves freeway traffic flow inference results. In fact, intercity freeway traffic is caused by socio-economic interactions across regions (Thompson et al., 2019), and the findings of this study support the accuracy of this claim in inferring freeway traffic. Although the existing study infers traffic volume from the perspective of formation, it only takes into account the impact of potential destination cities (Zhang et al., 2023), ignoring the spatial interaction processes between the cities of departure and destination. The experimental outcomes of this research show that the R^2 of the proposed ODBC-based method is improved by 0.2 %–23.3 % and the RMSE is reduced by 24 % compared to the weighted BC baseline model that only considers the influence of potential destination cities. Compared to the weighted BC baseline model that only considers the influence of potential departure cities, the proposed method improves R^2 by 25.8 % and reduces RMSE by 26 %. What’s more, the results of the proposed model for unknown road segments are better than those of the baseline models that solely consider the attractiveness of the destination or origin cities. This finding indicates that adequate consideration of the spatial interaction processes between the cities of departure and destination can better infer freeway traffic. In summary, this study provides a new scientific method for freeway traffic volume inference.

There are still some limitations in this research. First, the values of the proposed ODBC metrics may differ by using various distance functions. Future studies will thoroughly examine the effect of distance and fit the best distance function to increase the precision of inferring freeway traffic flow. Second, inferred indicators are obtained by integrating socio-economic indicators of origin and destination cities. However, it was unable to differentiate the specific contributions of origin and destination cities on freeway traffic flow. In order to solve this problem, future research will employ techniques like the causal graph attention model to clarify the contribution weights of origin and destination cities.

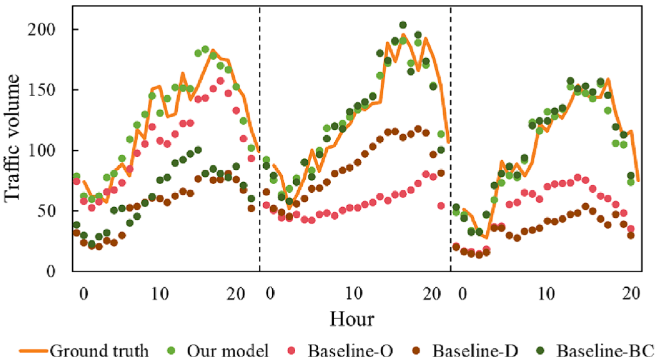

Fig. 7. Comparison of prediction results of unknown road segments.

5. Conclusion

In this study, a new freeway traffic volume inference method with city socio-economic interactions enhanced BC is proposed. The effectiveness of the proposed method is verified by employing traffic volume data from different types of vehicles collected at freeway toll stations in the Beijing-Tianjin-Hebei area of China. The findings indicate that the ODBC-based method for inferring freeway traffic flow outperforms both the BC-based baseline method and the BC weighting baseline methods relying solely on the attractiveness of potential departure or destination cities. Moreover, the proposed model exhibits superior accuracy and stability in predicting traffic volume for unknown road segments. In addition, SHAP is employed to quantify the effect vectors of the factors and the interaction between the factors on the freeway traffic volume. The results indicate that the primary industry ratio in both origin and destination cities is an important factor influencing freeway traffic volume. In summary, the proposed method is interpretable and enhance our comprehension of how city socio-economic interactions impact the flow of traffic on intercity freeways. This deeper insight serves to clarify the underlying driving mechanisms behind intercity freeway traffic.

CRediT authorship contribution statement

Beibei Zhang: Conceptualization, Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Shifen Cheng:** Supervision, Writing – review & editing. **Peixiao Wang:** Formal analysis. **Feng Lu:** Funding acquisition, Resources, 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

The authors do not have permission to share data.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [grant number 42371469 and 42101423]; Innovation Project of LREIS [grant number KPI003 and YPI002]; China National Postdoctoral Support Program for Innovative Talents [grant number

BX20230360]; China Postdoctoral Science Foundation [grant number 2023M743454]. We express our sincere gratitude for their support. We also express our gratitude to the anonymous referees for their invaluable remarks.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jag.2024.103818.

References

Breiman, L., 2001. Random forests. Mach Learn 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, J., Yang, L., Qin, C., Yang, Y., Peng, L., Ge, X., 2024. Heterogeneous graph traffic prediction considering spatial information around roads. Int. J. Appl. Earth Obs. Geoinformation 128, 103709. https://doi.org/10.1016/j.jag.2024.103709.

Cheng, S., Lu, F., Peng, P., Wu, S., 2018. Short-term traffic forecasting: an adaptive ST-KNN model that considers spatial heterogeneity. Comput. Environ. Urban Syst. 71, 186–198. https://doi.org/10.1016/j.compenvurbsys.2018.05.009.

Cheng, S., Lu, F., Peng, P., 2021. Short-term traffic forecasting by mining the non-stationarity of spatiotemporal patterns. IEEE Trans. Intell. Transp. Syst. 22, 6365–6383. https://doi.org/10.1109/TITS.2020.2991781.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232. http://www.jstor.org/stable/2699986.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. https://doi.org/10.1016/S0167-9473(01)00065-2.

Gao, S., Wang, Y., Gao, Y., Liu, Y., 2013. Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. Environ. Plan. B Plan. Des. 40, 135–153. https://doi.org/10.1068/b38141.

Henry, E., Bonnetain, L., Furno, A., Faouzi, N.-E.-E., Lyon, U., Zimeo, E., 2019. Spatio-temporal Correlations of betweenness centrality and traffic metrics. In: 2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). IEEE, pp. 1–10.

Kazerani, A., Winter, S., 2009. Can betweenness centrality explain traffic flow?. In: 12th AGILE International Conference on Geographic Information Science. Leibniz Universität Hannover, Germany, pp. 1–9.

Li, B., Gao, S., Liang, Y., Kang, Y., Prestby, T., Gao, Y., Xiao, R., 2020a. Estimation of regional economic development indicator from transportation network analytics. Sci. Rep. 10, 1–15. https://doi.org/10.1038/s41598-020-59505-2.

Li, F., Jia, H., Luo, Q., Li, Y., Yang, L., 2020b. Identification of critical links in a large-scale road network considering the traffic flow betweenness index. PLoS One 15. https://doi.org/10.1371/journal.pone.0227474.

Li, T., Wang, J., Huang, J., Yang, W., Chen, Z., 2021. Exploring the dynamic impacts of COVID-19 on intercity travel in China. J. Transp. Geogr. 95, 103153 https://doi.org/10.1016/j.jtrangeo.2021.103153.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30 https://doi.org/10.1016/j.ophtha.2018.11.016.

Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.W., Newman, S.F., Kim, J., Lee, S.I., 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. 2, 749–760. https://doi.org/10.1038/s41551-018-0304-0.

Pazoky, S.H., Pahlavani, P., 2021. Developing a multi-classifier system to classify OSM tags based on centrality parameters. Int. J. Appl. Earth Obs. Geoinformation 104, 102595. https://doi.org/10.1016/j.jag.2021.102595.

Petridis, N.E., Petridis, K., Stiakakis, E., 2020. Global e-waste trade network analysis. Resour. Conserv. Recycl. 158 https://doi.org/10.1016/j.resconrec.2020.104742.

Song, Y., Thatcher, D., Li, Q., McHugh, T., Wu, P., 2021. Developing sustainable road infrastructure performance indicators using a model-driven fuzzy spatial multicriteria decision making method. Renew. Sustain. Energy Rev. 138 https://doi.org/10.1016/j.rser.2020.110538.

Thompson, C.A., Saxberg, K., Lega, J., Tong, D., Brown, H.E., 2019. A cumulative gravity model for inter-urban spatial interaction at different scales. J. Transp. Geogr. 79, 102461 https://doi.org/10.1016/j.jtrangeo.2019.102461.

Turner, A., 2007. From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis. Environ. Plan. B Urban Anal. City Sci. 34, 539–555. https://doi.org/10.1068/b32067.

Wang, Y., Huang, Z., Yin, G., Li, H., Su, Y., Liu, Y., Shan, X., 2022b. Applying ollivier-ricci curvature to indicate the mismatch of travel demand and supply in urban transit network. Int. J. Appl. Earth Obs. Geoinformation 106, 102666. https://doi.org/10.1016/j.jag.2021.102666.

Wang, X., Pei, T., Song, C., Chen, J., Liu, Y., Guo, S., Chen, X., Shu, H., 2023b. X-index: a novel flow-based locational measure for quantifying centrality. Int. J. Appl. Earth Obs. Geoinformation 117, 103187. https://doi.org/10.1016/j.jag.2023.103187.

Wang, P., Zhang, T., Zheng, Y., Hu, T., 2022a. A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation. Int. J. Geogr. Inf. Sci. 36, 1231–1257. https://doi.org/10.1080/13658816.2022.2032081.

Wang, P., Zhang, Y., Hu, T., Zhang, T., 2023a. Urban traffic flow prediction: a dynamic temporal graph network considering missing values. Int. J. Geogr. Inf. Sci. 37, 885–912. https://doi.org/10.1080/13658816.2022.2146120.

Wen, T.H., Chin, W.C.B., Lai, P.C., 2017. Understanding the topological characteristics and flow complexity of urban traffic congestion. Phys. Stat. Mech. Its Appl. 473, 166–177. https://doi.org/10.1016/j.physa.2017.01.035.

Wu, X., Cao, W., Wang, J., Zhang, Y., Yang, W., Liu, Y., 2022a. A spatial interaction incorporated betweenness centrality measure. PLoS One 17, 1–20. https://doi.org/10.1371/journal.pone.0268203.

Wu, X., Yang, D., Wu, R., Gu, J., Wen, Y., Zhang, S., 2022b. High-resolution mapping of regional traffic emissions using land-use machine learning models. Atmos. Chem. Phys. 1939–1950 https://doi.org/10.5194/acp-22-1939-2022.

Yang, Y., Lu, X., Chen, J., Li, N., 2022. Factor mobility, transportation network and green economic growth of the urban agglomeration. Sci. Rep. 12, 1–12. https://doi.org/10.1038/s41598-022-24624-5.

Yang, Y., Jia, B., Yan, X.-Y., Chen, Y., Song, D., Zhi, D., Wang, Y., Gao, Z., 2023. Estimating intercity heavy truck mobility flows using the deep gravity framework. Transp. Res. Part E Logist. Transp. Rev. 179, 103320 https://doi.org/10.1016/j.tre.2023.103320.

Yi, Z., Liu, X.C., Markovic, N., Phillips, J., 2021. Inferencing hourly traffic volume using data-driven machine learning and graph theory. Comput. Environ. Urban Syst. 85 https://doi.org/10.1016/j.compenvurbsys.2020.101548.

Zhang, B., Cheng, S., Lu, F., Lei, M., 2022a. Estimation of exposure and premature mortality from near-roadway fine particulate matter concentrations emitted by heavy-duty diesel trucks in Beijing. Environ. Pollut. 311, 119990 https://doi.org/10.1016/j.envpol.2022.119990.

Zhang, B., Cheng, S., Zhao, Y., Lu, F., 2023. Inferring intercity freeway truck volume from the perspective of the potential destination city attractiveness. Sustain. Cities Soc. 98, 104834 https://doi.org/10.1016/j.scs.2023.104834.

Zhang, M., Huang, T., Guo, Z., He, Z., 2022b. Complex-network-based traffic network analysis and dynamics: a comprehensive review. Phys. Stat. Mech. Its Appl. 607, 128063 https://doi.org/10.1016/j.physa.2022.128063.

Zhao, Y., Cheng, S., Lu, F., 2023. Spatiotemporal interaction pattern of the Beijing agricultural product circulation. J. Geogr. Sci. 33, 1075–1094. https://doi.org/10.1007/s11442-023-2120-z.

Zhao, Y., Cheng, S., Zhang, B., Lu, F., 2024. Identifying the cargo types of road freight with semi-supervised trajectory semantic enhancement. Int. J. Geogr. Inf. Sci. 38, 432–453. https://doi.org/10.1080/13658816.2023.2288116.