Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 1191239, 11 pages
https://doi.org/10.1155/2022/1191239

Research Article

Traffic Risk Assessment Based on Warning Data

Tao Wang,[1] **Binbin Chen**,[1] **Yuzhi Chen**,[2] **Shejun Deng**,[3] **and Jun Chen**[2]

[1]*Guangxi Education Department Key Laboratory of ITS, Guilin University of Electronic Technology, Guilin 541004, China*
[2]*School of Transportation, Southeast University, Nanjing 211198, China*
[3]*College of Civil Science and Engineering, Yangzhou University, Yangzhou 225002, China*

Correspondence should be addressed to Tao Wang; wangtao@guet.edu.cn

Received 14 May 2022; Revised 29 July 2022; Accepted 1 August 2022; Published 27 August 2022

Academic Editor: Yajie Zou

To address the issues of insufficient danger excavation and long data collection period in traditional traffic risk assessment methods, this paper proposes a risk assessment method based on driver's improper driving behavior and abnormal vehicle state warning data. Meanwhile, this paper analyses the built environment's impact on traffic risk using the spatial econometric model. Firstly, a risk assessment system with the relative incidence of driver's improper driving behavior (eye closure, yawn, and looking away) and abnormal vehicle state (rapid acceleration, rapid deceleration, and lane departure) warnings as assessment indicators is constructed. Then, the risk responsibility weights of each warning type were determined using the entropy weight method. The risk classification thresholds were determined based on the Gaussian Mixture Model algorithm. Finally, a spatial econometric model was used to quantify the impact of built environment factors characterized by Point of Interest (POI) data on regional traffic risk, with the results of risk class classification as the dependent variable. The data of bus vehicle warnings in Zhenjiang, Jiangsu Province, are employed as an example for validation. The geographic cell of 1 km × 1 km scale is applied as the basic risk assessment unit. The results show that the optimal risk classification threshold for road traffic risk levels I and II is 1.92, the accuracy rate of class classification is 79.3%; the optimal risk classification threshold for levels II and III is 0.75, and the accuracy rate of class classification is 83.4%. The number of residential areas, Point of Interest (POI) mixing degree, and bus stops were significantly and positively correlated with transit traffic risk. The study results provide references for developing customized accident prevention measures and the appropriate setting of urban supporting facilities.

1. Introduction

Road traffic accident is one of the top ten causes of death around the world. According to the data published by the European Commission, there were approximately 18,800 deaths due to road accidents in the EU in 2020 [1]. In 2020, there are about 245,000 traffic accidents in China, with more than 61,700 deaths and 250,700 injuries [2]. Road traffic accidents pose a serious threat to human life and property safety. Reducing the risk of traffic accidents has become an urgent problem in the current transportation industry. According to Heinrich's pyramid theory, a major injury accident is not an isolated individual event but is usually accompanied by a large number of accident signs before a major injury accident occurs. Nascimento et al. [3] pointed out that Heinrich's safety pyramid relationship exists in many fields. If applied to road traffic accident prevention, it can be considered that each traffic accident occurs due to the accumulation of a large number of accident hazards. This provides an idea to reduce the occurrence of road traffic accidents, i.e., discovering more potential areas of traffic accidents through road traffic risk assessment to nip them in the bud.

At present, road traffic risk assessment can be classified according to data sources into (i) Risk assessment based on traffic accident data using methods such as accident number method, accident rate method, Bayesian, and Support Vector Machine (SVM). (ii) Risk assessment based on Surrogate Safety Measure (SSM) [4]. (iii) Starting from the four road traffic system elements of driver characteristics, vehicle conditions, road conditions, and environmental characteristics, traffic risk assessment is carried out using

Analytic Hierarchy Process (AHP) and Fuzzy Evaluation Method. (iv) Road traffic risk research is based on dangerous driving events, mainly including abnormal vehicle states represented by rapid acceleration and deceleration and driver's improper driving behaviors represented by eye closure, yawn, looking away, etc.

As for traffic accident data-based assessment, Cheng et al. [5] applied a Bayesian network to evaluate road traffic safety based on traffic accident data. Zhang et al. [6] selected traffic accident data and combined geographic information technology with systematic clustering to explore accident risk. It is worth noting that the source of traffic accident data is the traffic management department, while Benlagha and Charfeddine [7] conducted a related study based on accident data obtained by insurance companies. With the development of geographic information technology, GIS technology [8, 9] is also widely used in road safety analysis and assessment. The main reason is that traffic accidents are a spatial phenomenon. The spatial visualization function of GIS can be used to realize the visual display of accident location and risk area, while the traditional mathematical and statistical analysis methods primarily characterize the occurrence of accidents in the form of text and tables, which often ignore the spatial characteristics of traffic accidents. Also, since traffic accidents occur in both time and space dimensions, spatiotemporal risk studies using methods such as spatiotemporal network kernel density estimation and spatiotemporal cubes have emerged in recent years. Wang and Wang [10] explored the spatiotemporal characteristics of high-risk accident points based on the network spatiotemporal kernel density estimation method. Wu et al. [11] introduced the spatiotemporal cube into accident data mining, and the research results showed that the spatiotemporal cube method applies to traffic risk research at the meso-micro scale.

For road traffic risk assessment based on SSM, TTC and PET are common SSMs. SSM generally uses vehicle GPS, speedometers, video surveillance, radar sensors, and other onboard acquisition devices to quickly acquire the operational status of vehicles and other traffic participants. SSM can acquire a large amount of reliable data in a short period to identify high-risk locations before an accident occurs [4]. Yang et al. [12] fused accident data and SSM to construct a new operational safety assessment index to achieve the complementarity of both traffic accident data and SSM. For road safety assessment from four elements: driver characteristics, vehicle conditions, road conditions, and environmental characteristics, Yuan et al. [13] applied Extreme Gradient Boost-Apriori (XGB-Apriori) to explore the traffic accident risk of elderly pedestrians based on an in-depth analysis of the four elements. Liu et al. [14] established the urban road safety assessment system with index weights determined by the entropy weight method and finally validated the established index system and assessment model by taking Wuhan city as an example.

In terms of road safety risk research based on dangerous driving events, dangerous driving events are considered to be closely related to traffic risk [15], and Zhou and Zhang [16] pointed out that more than 90% of traffic accidents are related to dangerous driving events. Therefore, it is important to explore traffic risks from dangerous driving events. Currently, data on dangerous driving events are usually provided by onboard Diagnostics (OBD) or onboard recorders. However, in recent years, scholars have also been optimizing the identification methods of dangerous driving behaviors. Zhu et al. [17] proposed a risk index-based dangerous driving behaviour detection method using an unsupervised anomaly detection algorithm in video surveillance-based driving behaviour identification. Zou et al. [18] proposed a vehicle acceleration prediction model based on machine learning methods and driving behaviour analysis, which will be beneficial for the development of advanced driver assistance systems (ADAS) and the judgment of dangerous driving events. For risk assessment based on dangerous driving events, Cai et al. [19] extracted five types of dangerous driving events based on OBD data: rapid acceleration, rapid deceleration, rapid turning, overspeeding, and high-speed idling as assessment elements for road traffic risk assessment, and the results showed that the location of the most frequent dangerous driving incidents is roughly the same as the location of traffic accidents. Ren et al. [20] extracted four dangerous driving behaviors of speeding, rapid acceleration, rapid deceleration, and rapid turning using in-vehicle detection equipment for driving safety assessment, but there is a shortage of evaluation objects limited to individual drivers.

Relevant research in the field of traffic risk assessment has achieved rich achievements. However, there are still certain shortcomings: (i) Road traffic risk assessment based on traffic accident data has complicated procedures and difficulties in acquiring original traffic accident data, and at the same time, the acquired data have problems of insufficient accuracy, systematicity, and comprehensiveness [21]. Road safety studies usually require a certain magnitude of data. It results in a time-consuming collection of historical traffic accident data, which usually takes several years to obtain sufficient accident data. (ii) SSM-based studies are usually conducted within a limited number of locations, with insufficient regional scalability. (iii) The traffic risk assessment system established from four road traffic system elements, namely, driver characteristics, vehicle conditions, road conditions, and environmental characteristics, lacks consideration of driver behaviour and is ineffective in safety evaluation. (iv) Studies based on dangerous driving events usually consider only one type of abnormal vehicle state or driver's improper driving behaviour. In summary, there are not many studies using pre-accident sign warning data for road traffic risk assessment, and risk assessment based on warning data will uncover more hidden accident points and achieve pre-accident analysis.

The built environment refers to the manufactured environment established to meet the needs of human life, and Point of Interest (POI) data are usually applied to characterize the built environment [22]. Sallis et al. [23] pointed out that the built environment, represented by land use, is essential in causing urban traffic accidents. The scholar Wang and Xie [24] pointed out that the urban built environment has an important influence on traffic accidents, and traffic

conditions such as traffic flow and traffic speed are the main mediating factors linking the two. The existing studies tend to explore the influence of traffic conditions such as traffic flow and speed on traffic safety, but ignore that these traffic conditions intrinsically originate from the built environment. Some studies have used POI data for road traffic safety research. Jia et al. [25] used Kernel Density Estimation (KDE) and spatial clustering methods combined with POI data to explore the specific land factors associated with traffic accident distribution. Wang et al. [26] investigated the spatial effects of alcohol availability and DUI accidents using POI data. Mathew et al. [27] also included land-use characteristics in the assessment model when exploring the factors influencing adolescent traffic accidents. Similarly, most existing studies investigated the effects of the built environment on road safety based on traffic accident data.

This paper uses vehicle warning and POI data as the main data sources and combines the clustering method with spatial econometric models. The clustering method is applied to classify regional traffic risk levels. Finally, the spatial regression model is constructed based on the spatial heterogeneity of the urban built environment and traffic risk level. The influence of the urban built environment on regional road traffic risk is explored from a global perspective. Considering the current data availability, this paper uses the bus driver's improper driving behavior and abnormal vehicle state warning data.

2. Traffic Risk Assessment Methods

2.1. Calculating Risk Assessment Indicators. The lower the incidence of driver's improper driving behavior and abnormal vehicle state, the better the traffic safety; conversely, the higher the incidence of driver's improper driving behavior and abnormal vehicle state, the worse the traffic safety. This phenomenon is similar to the information entropy to describe the degree of the chaos of the system. The higher information entropy indicates a more chaotic system.

Therefore, with the help of information entropy, road traffic safety is expressed using the entropy value, known as the road traffic risk value. It is used as a primary evaluation index. The relative incidence of driver's improper driving behavior and abnormal vehicle state affecting traffic safety is used as a secondary evaluation index.

At the same time, the study area needs to be divided into geographical cells as the basic unit of safety assessment. So far, the calculation of road traffic risk values is divided into two steps. Firstly, the values of secondary evaluation indicators under various warning types are calculated. Then, the risk responsibility weights of secondary evaluation indicators are determined, along with the road traffic risk values.

The relative warning incidence of the secondary indicator P_{ij} is as follows:

$$P_{ij} = \frac{Z_{ij}}{\lambda_i}. \tag{1}$$

Here, i is the geographic cell number, j is the warning type number, Z_{ij} is the number of warning type j issued in the geographic cell i, and λ_i is the total number of warning vehicles in the geographic cell. The specific values of Z_{ij} and λ_i were obtained statistically using the spatial coordinate information of the warning points.

2.2. Calculating the Risk Responsibility Weights. Road traffic safety risks are not caused exclusively by a particular driver's improper driving behavior or abnormal vehicle state, and it is the result of the superposition of multiple risk factors [28]. However, the impact of each risk factor on road traffic safety cannot be the same. Therefore, this paper proposes the concept of risk responsibility weights to measure the impact of different factors on road traffic risk and achieve a comprehensive assessment of road traffic risk.

The risk responsibility weight is to determine the weight of different risk factors. The entropy weight method [29] is a typical method to determine the weight of indicators, and it is based on the variability of indicators to assess the objective weight. The smaller the degree of variability of a risk factor, the smaller the amount of information reflected, and the smaller the corresponding weight of indicators [30].

The entropy weight method is calculated as follows:

(1) *Data Standardization*. This paper selects the relative warning incidence P_{ij} within the cell as the assessment index. Because the higher the number of bus warnings in a cell, the higher the road traffic risk value, so the road traffic risk value is determined as a benefit indicator. The data standardization formula is as follows:

$$S_{ij} = \frac{\left(P_{ij} - \min P_{ij}\right)}{\max P_{ij} - \min P_{ij}}, \tag{2}$$

where i is the geographic cell number, and j is the warning type number.

(2) Calculation of index entropy value e_j.

$$e_j = -h \sum_{i=1}^{N} f_{ij} \ln f_{ij}, \tag{3}$$

where N is the total number of geographic cells, and $f_{ij} = S_{ij} / \sum_{i=1}^{N} S_{ij}$; $h = 1/\ln N$; Assume that when $f_{ij} = 0$, $f_{ij} \ln f_{ij} = 0$.

(3) Calculation of risk responsibility weights w_j.

$$w_j = \frac{1 - e_j}{\sum_{j=1}^{m} \left(1 - \mathrm{e}_j\right)}, \tag{4}$$

where m is the total number of warning types.

(4) Finally, the road traffic risk value R_i as

$$R_i = \sum_{j=1}^{m} P_{ij} w_j. \tag{5}$$

2.3. Determining the Optimal Risk Classification Number. Global spatial autocorrelation analysis can be applied to determine whether there are spatial aggregation

characteristics of driver's improper driving behaviors and abnormal vehicle state warnings. This is the basis for exploring road risks through clustering methods.

The *Moran's I* statistic is chosen to measure the global spatial autocorrelation. This paper takes the delineated geographic cell as the object of study, and *Moran's I* is calculated as

$$I = \frac{N}{\sum_a \sum_b w_{a,b}} \frac{\sum_a \sum_b w_{a,b} (R_a - \overline{R})(R_b - \overline{R})}{\sum_a (R_a - \overline{R})^2}, \quad (6)$$

where $w_{a,b}$ is the spatial weight between cell a and cell b; N denotes the total number of cells; R_a, R_b denote the road traffic risk values of the cell a, b; and $\overline{R}$ denotes the average value of all cells.

Moran's I range from [−1, 1]. If the I value > 0, it means that the spatial distribution of road traffic risk values of the cell is positively correlated, and closer to 1 means that its spatial aggregation is stronger. If the I value < 0, it means that the spatial distribution of road traffic risk values of the cell is negatively correlated. The closer to 1 means that its variability is greater, and the spatial distribution of road traffic risk values is more discrete. If the I value = 0, it means that it is not spatially correlated.

The significance test was performed with the statistical test value Z score, which was calculated as

$$Z(I) = \frac{I - E(I)}{\sqrt{\mathrm{VAR}(I)}}, \quad (7)$$

where $E(I)$ is the expected value of *Moran's I* and $\mathrm{VAR}(I)$ is the variance of *Moran's I*.

Lu and Cheng [31] pointed out that in terms of traffic risk assessment, a clear risk classification standard should be proposed after establishing a reasonable risk assessment index system. The road traffic risk classification is beneficial for the relevant departments to develop highly targeted road safety enhancement and improvement measures to obtain the highest safety return with the minimum investment. This paper uses the Gaussian Mixture Model (GMM) to select the optimal classification threshold. GMM is not only an unsupervised clustering algorithm but also a probability-based soft clustering algorithm, which has the advantages of a wide range of applicable shapes and insensitivity to noisy data in large datasets. The Expectation-Maximum (EM) algorithm is usually applied to solve the GMM in the study. Since the objective function of the EM algorithm is a nonconvex function, only the local optimum is guaranteed to be found. To avoid serious deviation of clustering results from the global optimum and ensure the speed of model convergence, K-means is used to initialize the parameters when using the GMM-EM algorithm [32]. The GMM algorithm is employed to implement bus risk clustering. The selected clustering indicator is a three-dimensional vector $\mathbf{x_i} \in \mathbf{R}^3$ composed of the latitude and longitude of the tuple center and the risk value, which contains the spatial characteristics of the geographic tuple and the transit warning characteristics within the tuple, and the vector structure is as follows:

$$\mathbf{x_i} = \left[x_{i,\mathrm{longitude}}\ x_{i,\mathrm{latitude}}\ R_i \right], \quad (8)$$

where $x_{i,\mathrm{longitude}}$, $x_{i,\mathrm{latitude}}$ are the latitude and longitude of the center of the cell i, and R_i is the road traffic risk value of the cell i. The clustering index needs to be normalized before calculation, and the attributes are scaled to between [0, 1].

In this paper, the optimal number of clusters of GMM is determined by using the Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) [33, 34], i.e., determining the number of risk classification levels, which is calculated as follows:

$$AIC = -2\ln(L) + 2c, \quad (9)$$

$$BIC = c\ln(n) - 2\ln(L), \quad (10)$$

where L is the objective function of the EM algorithm, n is the number of samples, and c is the number of degrees of freedom in the GMM. From equations (9) and (10), it can be seen that the smaller the value of *AIC* and *BIC*, the better the application of the Gaussian Mixture Model, and the minimum value at this time is also the best number of clusters.

Compared with AIC, BIC penalizes the model parameters more when the data volume is large, making it easier to choose a model with a small number of clusters and effectively avoid dimensional catastrophe. Therefore, when selecting the optimal number of clusters, the *BIC* value is preferred to choose the optimal number of clusters.

2.4. Calculating the Risk Level Threshold. Road traffic risk classification is conducive for prioritizing regional traffic risk control and determining graded control measures. If the optimal number of risk levels determined is K, then $K-1$ level classification thresholds need to be determined. Taking the example of dividing two classes, the steps for solving the class division threshold are as follows:

Step 1: count the total number of cells clustered as category 1 and category 2, denoted as l, and rank them from smallest to largest according to the value at risk.

Step 2: assuming that the initial risk value threshold for category 1 and 2 classification s_1 is =0, the threshold value s_1 is used to classify road traffic risk values category 1 and 2 into two classes.

Step 3: count the number of cells clustered as class 1, but classified as class 2 by the grading threshold, denoted as q_1; the number of cells clustered as class 2, but classified as class 1 by the grading threshold, denoted as q_2.

Step 4: calculate the accuracy of the division threshold [19].

$$\rho = 1 - \frac{q_1 + q_2}{l} \times 100\%. \quad (11)$$

Step 5: increase the grading threshold by 0.01 until the data point with the largest risk value is covered, calculate in turn, and select the grading threshold with the highest accuracy as the grading threshold for road traffic risk classes 1 and 2.

3. Example Analysis

3.1. Study Data. The bus warning data were obtained from the Bus Driver State Monitoring System (DSMS) and Advanced Driver Assistance System (ADAS) in Zhenjiang City. The data information mainly contains four aspects: driver information, vehicle identification information, warning situation information, and equipment identification information, and the data period is January 8, 2019–January 12, 2019. The data relate to three administrative districts of Zhenjiang, namely, Runzhou District, Dantu District, and Jingkou District. The warnings include driver's improper driving behavior (eye closure, yawn, looking away) and abnormal vehicle state warnings (lane departure, rapid acceleration, rapid deceleration). Since the data collection is often affected by various external factors, there are some missing and incorrect data, so the original data were screened and cleaned to obtain 23538 pieces of valid bus warning data involving 390 bus vehicles. The study mainly involves the spatial distribution of bus warnings, so the data shown in Table 1 are extracted for the subsequent study, including five attributes: vehicle code, warning time, warning type, and latitude and longitude of the warning.

In consideration of the effect of the follow-up study, the study area was cut along the Shanghai-Chengdu Expressway in Dantu District based on the existing administrative division, and the reserved study area is shown in Figure 1. At the same time, only the public transport warning data within the study area were retained. In the selection of the evaluation analysis unit, the cell size of 1 km × 1 km is used as the basic unit of road traffic risk assessment, referring to the common cell size of traffic accident analysis. The study area was divided into a total of 1419 geographic cells, and the results are shown in Figure 2.

3.2. Spatial Distribution Characteristics. As shown in Figure 3, the Kernel Density Estimation (KDE) was used to analyze the spatial distribution of the warning data by time. It can be found that the warning high-density areas show a multi-core point spatial distribution pattern during 0:00–6:00 hours, mainly in the two downtown areas of Runzhou and Jingkou District and the roads near Zhenjiang South Station. During the 6:00–24:00 hours, the warning high-density area shows the characteristics of strip distribution along the road, mainly involving the roads of Zhongshan West Road, Zhongshan East Road, Xuefu Road, and Dingmao Bridge Road in the city center. The density distribution shows an overall trend of decreasing spread from the city center to the surrounding area.

3.3. Road Traffic Risk Assessment. To obtain the traffic risk status of different geographic cells, the warning data and the geographic cell division data are spatially matched. Then, the number of each warning type of the geographic metacell can be acquired, and thus the safety assessment index P_{ij} is derived from equation (1). The risk responsibility weights w_j of different warning types are calculated according to equation (4), as shown in Table 2. As can be seen from Table 2, the risk responsibility weights of the warning types within the study area are in the order of looking away, lane departure, rapid acceleration, rapid deceleration, yawn, and eye closure. It can also be found that the risk responsibility weights of driver's improper driving behavior warning and abnormal vehicle state warning are basically equal. They are 0.528 and 0.472, respectively. The road traffic risk value of each cell is finally determined by combining it with equation (5). The distribution of road traffic risk values of nonzero risk value cells is shown in Figure 4.

The spatial autocorrelation results showed that the *Moran's I* index of road traffic risk values for cells was 0.60 (>0) and 31.72 ($p < 0.01$). It is known that the possibility of randomly generated clusters of cells within the study area is less than 1% and is statistically significant, so it is reasonable to explore the bus road traffic risk by the clustering method.

The result of calculating the optimal number of class divisions is shown in Figure 5. When the number of clusters is 3, the BIC curve reaches the minimum value. That is, the optimal number of clusters is 3. Thus, the road traffic risk is divided into three classes. The distribution characteristics of road risk values for the three categories are shown in Figure 6. The calculated average road traffic risk values of the three categories are 0.42, 1.09, and 1.90, respectively. It can be inferred that the overall road traffic risk values of the three categories have the relationship of category 1 < category 2 < category 3, which lays the foundation for determining the subsequent class classification thresholds.

As shown in the previous section, it is necessary to calculate the optimal grading thresholds between category 1 and category 2, and between category 2 and category 3. In order to determine the best classification threshold, the road traffic risk values are arranged in order from smallest to largest, as shown in Figure 7(a). From the calculation of equation (11), it can be obtained that when the road traffic risk threshold is 0.75, the classification accuracy of category 1 and category 2 reaches a maximum of 83.4%. When the road traffic risk threshold is 1.92, the class classification accuracy of category 2 and category 3 reaches a maximum of 79.3%, and the result is shown in Figure 7(b). Therefore, the regional road traffic risk can be defined as low (III), medium (II), and high (I) levels. The risk value of the low (III) risk region is taken as (0, 0.75), the medium (II) risk region is taken as (0.75, 1.92), and the high (I) risk region is taken as (1.92, 5.87), where 5.87 is the maximum risk value of the cell in this risk assessment obtained in the previous section.

3.4. Analysis of Assessment Results. The results of risk ungraded space display of geographic cells are shown in Figure 8. The darker the color, the higher the risk value of the cell. It can be found that when the road traffic risk is ungraded, the risk value calculation results are not effective for regional risk identification, which is not conducive for the later development of targeted safety improvement measures. The results of regional risk classification are shown in Figure 9. Considering that the cells with a risk value of 0 are mostly areas without bus routes, the area with a risk value of 0 is marked out separately. Then, the risk response level is defined for cells with risk values greater than 0, and the risk level is defined as high risk (I), medium risk (II), and low risk (III) levels.

Table 1: Bus warning data.

Property name	Vehicle code	Warning time	Warning type	Longitude (°)	Latitude (°)
Example	18200215	2019/01/08 05:16:34	Eye closure	119.460232	32.159400

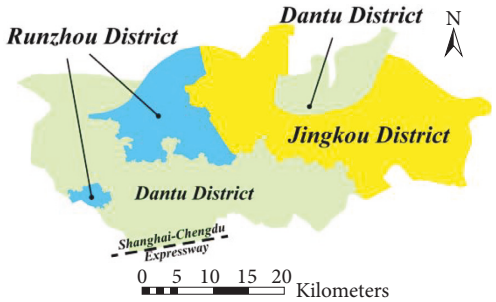

Figure 1: Area of research.

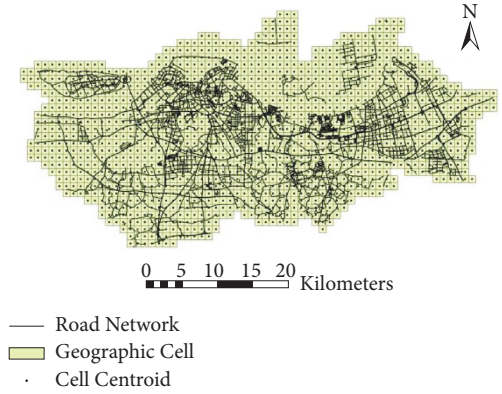

Figure 2: Geographic cell division.

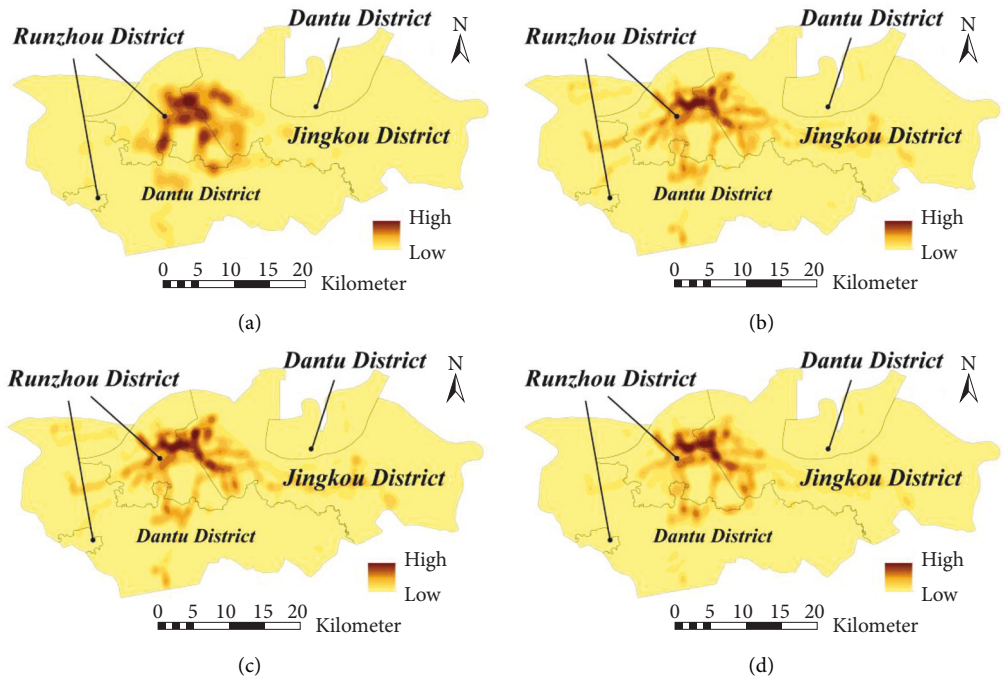

Figure 3: Kernel density by the period. (a) 0:00–6:00. (b) 6:00–12:00. (c) 12:00–18:00. (d) 18:00–24:00.

Table 2: Risk liability weight.

Indicators	Driver's improper driving behaviours			Abnormal vehicle states		
	Eye closure	Yawn	Looking away	Rapid acceleration	Rapid deceleration	Lane departure
w_j	0.119	0.126	0.283	0.156	0.135	0.181

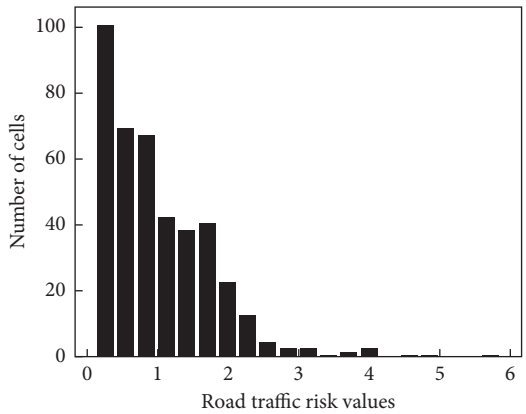

Figure 4: Road traffic risk value distribution.

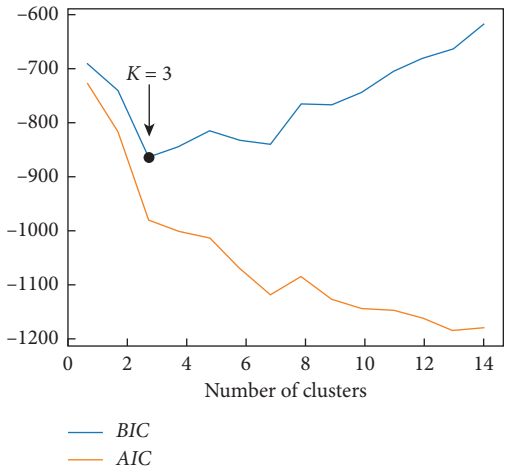

Figure 5: Selection of the best number of clusters.

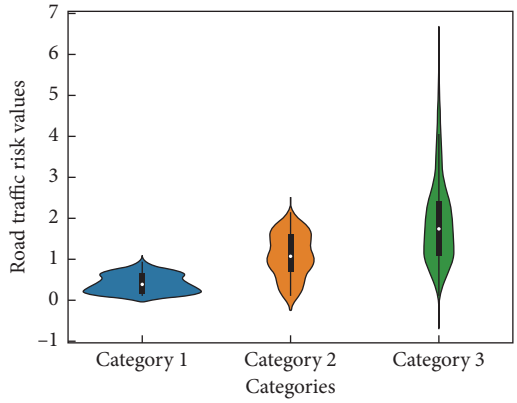

Figure 6: Category risk distribution characteristics.

Comparing Figure 9 with the kernel density estimation results in Figure 3, it can be seen that there are differences in the high-risk regions. For example, in the kernel density estimation results of Figure 3 for each period, the color of the edge areas in the Jingkou district are all relatively light, indicating that the risk level of these areas is not high when considered from the perspective of warning frequency alone. However, according to this paper, the road traffic risk classification results show that areas such as Zhenjiang New District Management Committee and Yinshan Xincheng at the edge of Jingkou District are also marked with dark color as high-risk areas, as shown in Figure 9. Analogous to the traffic accident black spot judgment, if the regional traffic volume is high, the high frequency of traffic accidents at this time does not indicate that the area is an accident black spot. Instead, the ratio of the number of traffic accidents at a location to the corresponding average daily traffic flow is usually used as an indicator for judging road accident black spots rather than from the perspective of accident frequency alone. Therefore, the kernel density estimation method considered from the perspective of warning quantity alone does not fully reflect the regional traffic risk situation.

3.5. Traffic Risks in Relation to the Built Environment. The regression estimation of spatial econometric models can elucidate the spatial dependence between urban built environment and traffic safety risks. In this paper, we use the AutoNavi Map API to obtain POI data of Zhenjiang City, with a total of 194,994 items in 23 major categories. Due to the redundancy of the original POI data, this paper reclassifies the original POI data into six categories of residential land, public services, commercial services, industrial land, transportation facilities, and green squares according to the Code for Classification of Urban Land Use and Planning Standards of Development Land (GB50137-2011), and eliminates the points with low public recognition [35]. Considering that this paper uses transit warning data, the bus stops were separately extracted from the transportation facility types.

The commonly used regression models are the Ordinary Least Squares (OLS) model, Spatial Lag Model (SLM), and Spatial Error Model (SEM). The OLS is no longer applicable when there is an obvious spatial dependence of the model residuals, and the spatial weight matrix needs to be introduced, i.e., SLM, SEM, and other spatial econometric models,where the primary expression of the SLM is as follows:

$$y = \theta W y + X\beta + \varepsilon, \tag{12}$$

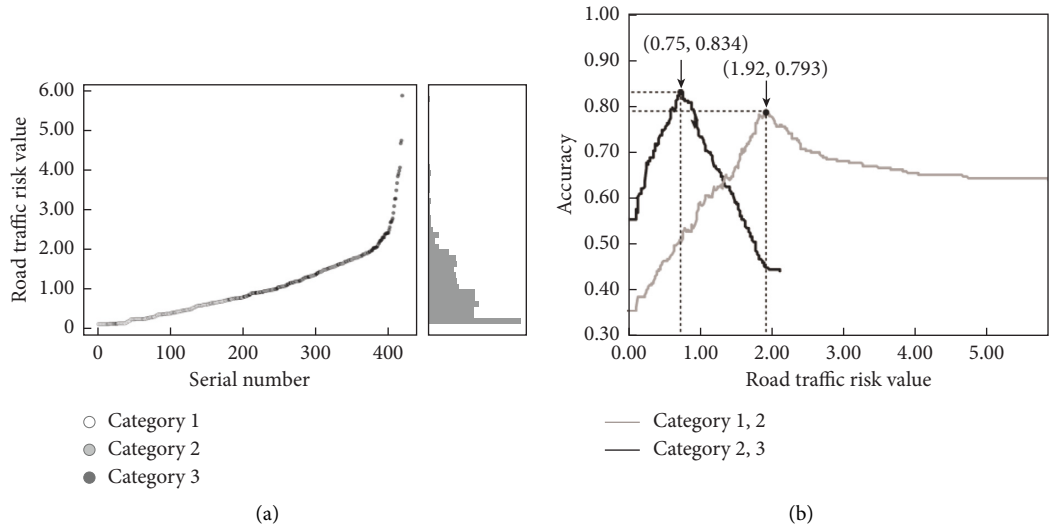

Figure 7: Determination of the classification threshold. (a) Road traffic risk value ranking. (b) Classification thresholds.

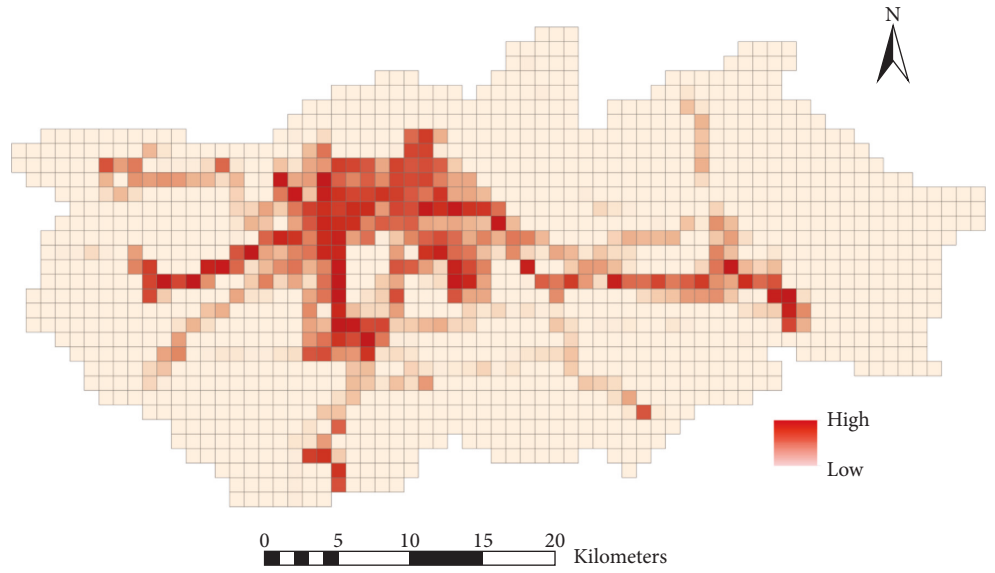

Figure 8: Road traffic risks.

where y is the dependent variable. X is the independent variable. W is the spatial weight matrix. Wy is the spatial lagged variable. ε is the error term. θ is the spatial autoregressive coefficient to be estimated. β is the coefficient of the independent variable to be estimated.

Anselin [36] recommended a selection process for selecting the appropriate spatial econometric model based on OLS regression with Lagrange Multiplier (LM) test and Robust LM.

This paper constructs a regression model using the regional risk level as the dependent variable to explore the influence of the built environment on regional traffic risk. As shown in Table 3, based on the number of POIs in the geographic cell as the explanatory variables, two indicators, total POI distribution and POI mixing degree [37], are introduced to characterize the concentration and functional diversity of building facilities in the analysis unit. The model was constructed by stepwise regression analysis of the explanatory variables to be selected, and the three variables "Residential," "POI_gini," and "Bus stop" were retained. The Variance Inflation Factor (VIF) of these three variables was less than 5 by the multicollinearity diagnostics, which means they passed the collinearity diagnostics, so they were introduced into the model.

In this paper, the spatial weights are constructed using Euclidean distance. The residuals from the regression of the explanatory variables into the OLS model are subjected to *Moran's I* test. The test value is 0.198, and the p value is 0.000, which means that the original hypothesis of "spatial autocorrelation" is rejected at the 1% significance level. Therefore, the spatial dependence factor in the residuals of the

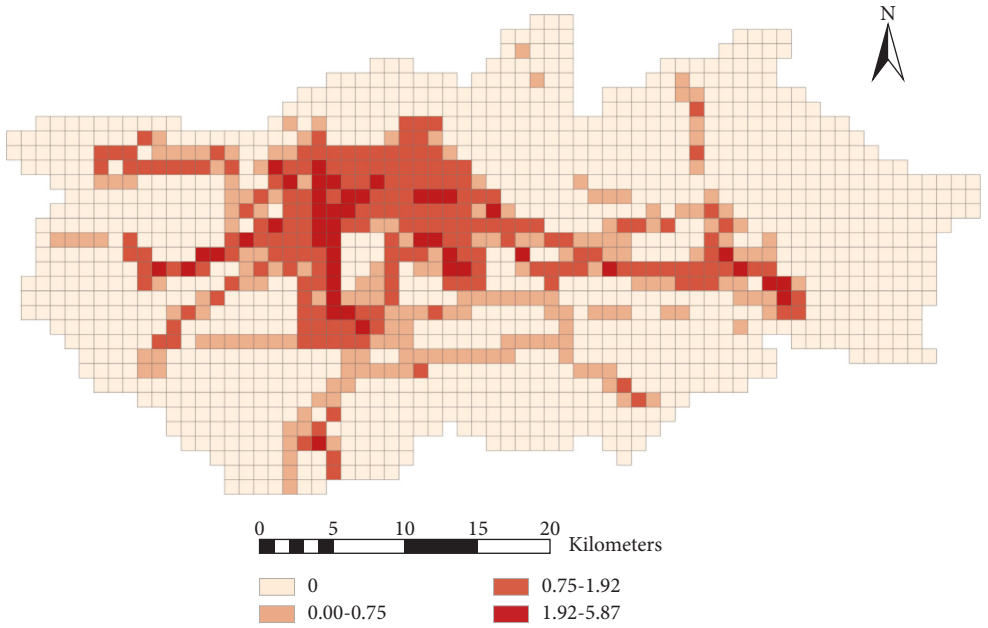

Figure 9: Road traffic risk classification.

Table 3: The explanatory variables to be selected for the model.

Explanatory variables to be selected		Mean value	Remarks
Built environment	Number of residential land-based POIs	2.84	Residential
	Number of public service-type POIs	19.00	Public
	Number of commercial service-oriented POIs	58.39	Commercial
	Number of industrial land-type POIs	11.67	Industrial
	Number of transportation facility-based POIs	5.94	Transportation
	Number of greenfield plaza-type POIs	1.42	Square
	Total POI distribution	99.27	POI_total
	POI mixing degree	2.22	POI_gini
	Number of bus stops	2.34	Bus stop

Table 4: Model check results.

Indicators	LM-lag	Robust LM-Lag	LM-error	Robust LM-error
Statistical quantities	48.212	8.818	39.415	0.021
p value	0.000	0.003	0.000	0.884

Table 5: Estimation results.

Variables	OLS	SLM	SEM
Constant	1.139^{***}	0.590^{***}	1.229^{***}
Residential	0.032^{***}	0.017^{**}	0.024^{***}
POI_gini	0.099^{***}	0.060^{*}	0.070^{**}
Bus stop	0.093^{***}	0.084^{***}	0.089^{***}
AIC	735.515	691.099	696.246
SC	751.657	711.277	712.387
Log-likelihood value	−363.758	−340.550	−344.123

Note: "*," "**," and "***" indicate significant effects at 0.10, 0.05, and 0.01 levels, respectively.

OLS model needs to be removed. The results of the Lagrange multiplier test are shown in Table 4, and the LM-Lag and LM-Error passed the significance test. The p values of Robust LM-Lag and Robust LM-Error statistics are 0.003 and 0.884, respectively; so the SLM is chosen for the causal investigation.

For the comparison of model selection effects, OLS, SLM, and SEM regressions were estimated using Geoda [38] software, and the results are shown in Table 5.

The log-likelihood value, AIC, and SC are important indicators for judging the merits of the model. As shown in Table 5, the log-likelihood value of the SLM model is −340.55, which is greater than OLS (−363.758) and SEM (−344.123). The AIC and SC of the SLM are 691.099 and 711.277, respectively, both of which are smaller than the corresponding indicators of OLS and SEM, indicating that the spatial lag model fits better than OLS and SEM. This further verifies the correctness of the LM test results.

The regression coefficients of residential, POI_gini, and bus stops in the SLM model are 0.017, 0.060, and 0.084, respectively. Meanwhile, the regression coefficients were all positive, indicating that the number of residential areas, POI mixing degree, and bus stops were significantly and positively related to bus safety risk. The greater the number of residential areas and the more complex the diversity of land functions, the more susceptible the bus vehicles are to pedestrians, vehicles, and intertwined roads during the operation of these areas, leading to an increase in risk. The number of bus stops is an important factor influencing the risk to bus safety. More bus stops will increase the number of bus stops per unit of time, and the driving speed needs to change frequently, leading to an increase in the probability of risk. This is supported by the study of Quddus [39], which states that speed change is positively correlated with accident rate, with a 1% increase in speed change associated with a 0.3% increase in the accident rate.

4. Conclusion

This paper establishes a new traffic risk assessment method based on driver's improper driving behavior and abnormal vehicle state warning data. The method remedies the problems of difficult sample data collection and insufficient scalability of traditional traffic safety assessment. Finally, this paper also innovatively uses the warning data to explore the spatial heterogeneity influence of urban built environment on regional road traffic risk.

(1) The road traffic risks in the study area should be classified into three classes. The best classification threshold for levels II and III is 0.75, and the classification accuracy rate is 83.4%. The best classification threshold for levels I and II is 1.92, and the classification accuracy rate is 79.3%. The research results lay the foundation for road traffic risk identification, regional safety refinement management, and targeted accident prevention countermeasures.

(2) Spatial Lag Model (SLM) has the best effect among the three econometric models. The model results show that the number of residential areas, POI mixing degree, and bus stops significantly positively affect regional transit road traffic risk.

This paper mainly assesses road traffic safety through the incidence of warning, while the actual operating environment and other characteristics will have a greater impact on vehicle operation, and subsequently consider the integration of multiple factors and establish a multi-dimensional data fusion assessment method. Meanwhile, this paper is a study using bus warning data, so only the road traffic risk assessment of bus vehicles is achieved. In the future, multi-vehicle risk warning data in an intelligent networked vehicle environment will support a more comprehensive road traffic risk assessment.

Data Availability

The raw data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Fund for Less Developed Regions of the National Natural Science Foundation of China (71861006, 61963011), Guangxi Natural Science Foundation (2020GXNSFAA159153), Guangxi Science and Technology Base and Talent Special Project (AD20159035), The Ministry of Education of Humanities and Social Science Project (19YJAZH011), and Guilin Key Research and Development Project (20210214-1).

References

[1] Directorate-General for Mobility and Transport, "Road safety: European Commission rewards effective initiatives and publishes 2020 figures on road fatalities," 2021, https://transport.ec.europa.eu/news/road-safety-european-commission-rewards-effective-initiatives-and-publishes-2020-figures-road-2021-11-18_en.

[2] "National bureau of statistics of the people's Republic of China, China statistical yearbook 2021," *China Statistical Yearbook 2021*, China's Statistics Press, Beijing, China, 2021.

[3] F. A. C. Nascimento, A. Majumdar, and W. Y. Ochieng, "Investigating the truth of Heinrich's pyramid in offshore helicopter transportation," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2336, no. 1, pp. 105–116, 2013.

[4] K. Xie, D. Yang, K. Ozbay, and H. Yang, "Use of real-world connected vehicle data in identifying high-risk locations based on a new surrogate safety measure," *Accident Analysis & Prevention*, vol. 125, pp. 311–319, 2019.

[5] W. Cheng, G. S. Gill, Y. P. Zhang, and Z. Cao, "Bayesian spatiotemporal crash frequency models with mixture components for space-time interactions," *Accident Analysis & Prevention*, vol. 112, pp. 84–93, 2018.

[6] G. N. Zhang, Q. T. Zhong, and Q. X. Yang, "Temporal-spatial characteristics and influencing factors of at-fault traffic crashes: a case study in Guangzhou," *Journal of Transportation Systems Engineering and Information Technology*, vol. 19, pp. 208–214, 2019.

[7] N. Benlagha and L. Charfeddine, "Risk factors of road accident severity and the development of a new system for prevention: new insights from China," *Accident Analysis & Prevention*, vol. 136, Article ID 105411, 2020.

[8] F. Ouni and M. Belloumi, "Pattern of road traffic crash hot zones versus probable hot zones in Tunisia: a geospatial analysis," *Accident Analysis & Prevention*, vol. 128, pp. 185–196, 2019.

[9] A. F. Ramírez and C. Valencia, "Spatiotemporal correlation study of traffic accidents with fatalities and injuries in Bogota (Colombia)," *Accident Analysis & Prevention*, vol. 149, Article ID 105848, 2021.

[10] Y. Z. Wang and L. J. Wang, "An identification method of traffic accident black point based on street-network spatial-temporal kernel density estimation," *Scientia Geographica Sinica*, vol. 39, pp. 1238–1245, 2019.

[11] P. J. Wu, X. H. Meng, and M. D. Cao, "Identification of black spots in urban roads and spatiotemporal patterns mining," *China Safety Science Journal*, vol. 30, pp. 127–133, 2020.

[12] D. Yang, K. Xie, K. Ozbay, and H. Yang, "Fusing crash data and surrogate safety measures for safety assessment: development of a structural equation model with conditional autoregressive spatial effect and random parameters," *Accident Analysis & Prevention*, vol. 152, Article ID 105971, 2021.

[13] Z. Z. Yuan, M. Z. Guo, Y. X. Peng, and Y. Yang, "Risk recognition of older pedestrian traffic crashes based on XGB-Apriori algorithm," *Journal of Transportation Systems Engineering and Information Technology*, vol. 22, pp. 195–208, 2022.

[14] S. Y. Liu, Y. H. Song, and D. Liu, "Risk assessment research of urban road traffic safety based on Extension Matter Element Model," *Journal of Physics: Conference Series*, vol. 1486, no. 7, Article ID 072020, 2020.

[15] X. X. Zhang, X. S. Wang, Y. Ma, and Q. B. Ma, "International research progress on driving behavior and driving risks," *China Journal of Highway and Transport*, vol. 33, pp. 1–17, 2020.

[16] T. Q. Zhou and J. Y. Zhang, "Analysis of commercial truck drivers' potentially dangerous driving behaviors based on 11-month digital tachograph data and multilevel modeling approach," *Accident Analysis & Prevention*, vol. 132, Article ID 105256, 2019.

[17] S. X. Zhu, C. Y. Li, K. X. Fang, Y. C. Peng, Y. M. Jiang, and Y. J. Zou, "An optimized algorithm for dangerous driving behavior identification based on unbalanced data," *Electronics*, vol. 11, no. 10, Article ID 1557, 2022.

[18] Y. J. Zou, L. S. Ding, H. Zhang, T. Zhu, and L. T. Wu, "Vehicle acceleration prediction based on machine learning models and driving behavior analysis," *Applied Sciences*, vol. 12, no. 10, Article ID 5259, 2022.

[19] X. Y. Cai, C. L. Lei, B. Peng, X. Y. Tang, and Z. G. Gao, "Road traffic safety risk estimation based on driving behavior and information entropy," *China Journal of Highway and Transport*, vol. 33, pp. 190–201, 2020.

[20] H. J. Ren, T. Xu, and X. Li, "Driving behavior analysis based on trajectory data collected with vehicle-mounted GPS receivers," *Geomatics and Information Science of Wuhan University*, vol. 39, pp. 739–744, 2014.

[21] J. B. Hu, L. C. He, and R. H. Wang, "Review of safety evaluation of freeway interchange," *China Journal of Highway and Transport*, vol. 33, pp. 17–28, 2020.

[22] B. Wang, Y. Q. Lei, C. G. Wang, and L. Wang, "The spatio-temporal impacts of the built environment on urban vitality: a study based on big data," *Scientia Geographica Sinica*, vol. 42, pp. 274–283, 2022.

[23] J. F. Sallis, F. Bull, R. Burdett et al., "Use of science to guide city planning policy and practice: how to achieve healthy and sustainable future cities," *The Lancet*, vol. 388, no. 10062, pp. 2936–2947, 2016.

[24] C. Wang and B. Xie, "Research progress on the driving mechanism of traffic accidents from the perspective of land use," *Progress in Geography*, vol. 39, no. 9, pp. 1597–1606, 2020.

[25] R. Jia, A. Khadka, and I. Kim, "Traffic crash analysis with point-of-interest spatial clustering," *Accident Analysis & Prevention*, vol. 121, pp. 223–230, 2018.

[26] S. H. Wang, Y. Y. Chen, J. L. Huang, Z. Liu, and J. Li, "Spatial effects of alcohol availability on drunk driving traffic accidents," *Journal of Beijing University of Technology*, vol. 45, pp. 886–894, 2019.

[27] S. Mathew, S. S. Pulugurtha, and S. Duvvuri, "Exploring the effect of road network, demographic, and land use characteristics on teen crash frequency using geographically weighted negative binomial regression," *Accident Analysis & Prevention*, vol. 168, Article ID 106615, 2022.

[28] T. Wang, Y. Z. Chen, X. C. Yan, J. Chen, and W. Y. Li, "The relationship between bus drivers' improper driving behaviors and abnormal vehicle states based on advanced driver assistance systems in naturalistic driving," *Mathematical Problems in Engineering*, vol. 2020, Article ID 9743504, 2020.

[29] F. F. Yuan, "A prioritizing method for railway passenger boarding scheme based on information entropy and cosine decision," *Railway Transport and Economy*, vol. 41, pp. 115–120, 2019.

[30] T. Wang, Y. Z. Chen, X. C. Yan, W. Y. Li, and D. Shi, "Assessment of drivers' comprehensive driving capability under man-computer cooperative driving conditions," *IEEE Access*, vol. 8, pp. 152909–152923, 2020.

[31] J. Lu and Z. Y. Cheng, "Research and development of road traffic network security risk identification," *Journal of Southeast University (Natural Science Edition)*, vol. 49, pp. 404–412, 2019.

[32] T. Liu, R. Fu, Y. Ma, Z. F. Liu, and W. D. Chen, "Car-following warning rules considering driving styles," *China Journal of Highway and Transport*, vol. 33, pp. 170–180, 2020.

[33] D. Liu, H. Chen, T. Li, H. Zhang, and Y. Geng, "Spatio-temporal differentiation of village ecosystem service bundles in the loess hilly and gully region and terrain gradient analysis," *Progress in Geography*, vol. 41, no. 4, pp. 670–681, 2022.

[34] L. Liu, W. Shi, X. P. Zhang, B. Han, X. Dong, and L. Yuan, "Research on spatial distribution of artificial fill in Xi'an based on Gaussian Mixture clustering algorithm," *Northwestern Geology*, vol. 55, pp. 298–304, 2022.

[35] J. Chi, L. M. Jiao, T. Dong, Y. Y. Guo, and Y. L. Ma, "Quantitative identification and visualization of urban functional area based on POI data," *Journal of Geomatics*, vol. 41, pp. 68–73, 2016.

[36] L. Anselin, *Spatial Econometrics: Methods and Models*, Kluwer Academic Publisher, Dordrecht, The Netherlands, 1988.

[37] E. Z. Yu and J. B. Zhou, "Travel characteristics and influencing factors of bike sharing based on Spatial Lag Model," *Journal of Transport Information and Safety*, vol. 39, pp. 103–110, 2021.

[38] L. Anselin, I. Syabri, and Y. Kho, "GeoDa: an introduction to spatial data analysis," *Geographical Analysis*, vol. 38, no. 1, pp. 5–22, 2006.

[39] M. Quddus, "Exploring the relationship between average speed, speed variation, and accident rates using Spatial Statistical Models and GIS," *Journal of Transportation Safety & Security*, vol. 5, no. 1, pp. 27–45, 2013.