Identification of potential traffic accident hot spots based on accident data and GIS

Hongge Zhu[1,a] Yuntong Zhou[2] Yanyan Chen[3]

[1] *Hongge Zhu, China Highway Engineering Consulting Corporation, Beijing University of Technology, China*
[2] *Yuntong Zhou, Beijing Engineering Research Center of Urban Transport Operation Guarantee; Beijing University of Technology Beijing, China*
[3] *Yanyan Chen, Beijing Key Laboratory of Traffic Engineering, Beijing University of Technology, Beijing, China*

Abstract. The problem of road traffic safety has been widely concerned in recent years. The identification of traffic accident hot spots can effectively improve the road traffic safety and let the traffic managers formulate targeted improvement measures and suggestions. The traditional identification method of accident hot spot does not consider the spatial attribute of the accident, so it has some limitations in the identification of traffic accident hot area. Therefore, this paper first proposes a method to identify the hot spot of traffic accidents based on geographic information system (GIS). The mathematical model and machine learning model are used to explore the correlation between traffic accidents and spatial characteristics from macro and micro aspects. Finally, taking Beijing as an example, the feasibility of the research method is proved by using the accident data of Beijing in 2015 and the geographic information of Beijing. The research results of this paper can realize the spatial effective transformation of accident records, comprehensively consider the micro and macro attributes of the accident itself, realize the automatic and efficient identification of the accident hot spot. In addition, the causality analysis results between each attribute and the distribution of accident hot spots can help decision makers to formulate safety and sustainable road strategies.

1 Introduction

China's traffic safety situation is grim. In recent years, the number of traffic accidents in China is increasing year by year, and the number of accident deaths ranks first in the world. Therefore, it is very important to identify potential traffic accident hotspots to ensure safe and smooth travel environment. The traditional traffic accident analysis usually relies on the map, engineering drawing and general information management system which lack of effective spatial data analysis, which results in not only the heavy workload of data processing, but also the analysis results are not comprehensive and accurate, and it is difficult to reveal the internal laws. Therefore, it is necessary to make full use of the information, technology and scientific methods to effectively analyze and identify the dangerous road sections, so as to find out the current dangerous areas of traffic accidents, and take measures such as identification, rectification and prevention. The occurrence of urban traffic accidents has clear spatial location characteristics, so it is necessary to analyze the relationship between traffic accidents and geographical location in the process of identifying traffic accident hot areas. Geographic Information System (GIS) is an important technical means to analyze and explore the problems related to traffic accidents and geographical location, and has the spatial and spatial database management function of visual interface.[1][2][3] [4]

The research significance of the text is as follows:

(1) This paper uses ArcGIS software to study. ArcGIS is a complete GIS software system based on industrial standards. It has the characteristics of comprehensive function, good scalability and user-defined flexibility. It can combine the traditional traffic accident database with the visualization and spatial analysis ability of GIS system, and use the traditional mathematical model algorithm and machine learning algorithm to find out the various attributes hidden behind the traffic accident data. The potential connection of data reveals the hidden regularity.

(2) The effective identification of potential accident hotspots is effective and practical for management decision-making, reducing traffic accidents and improving safety operation management.

[a] Corresponding author: 2047768025@qq.com

(3) It provides certain reference for traffic management departments to make decisions, which is of great significance to reduce the incidence of traffic accidents, reduce potential accident risks and improve the effectiveness and practicability of safety operation management.

2 Method

The purpose of this paper is to find out the relationship between the severity of vehicle accidents and macro and micro factors, and to provide the basis for the identification of traffic accident hot areas.

2.1 Binary logic regression

Based on the accident data, this study uses binary logistic regression to establish a model to analyze the impact of different road facilities on the severity of traffic accidents from the micro level. In the logistic regression model, the results are divided into two parts. The relationship between probability and events is described by the following link function:

$$P_n = \frac{e^{f(x_n)}}{1+e^{f(x_n)}} = \frac{1}{1+e^{-f(x_n)}} \quad (1)$$

Where, Is the probability of occurrence of the event; It's a linear function that explains variables.

In logistic regression model, the linear function is related to the expected value of response, which is composed of k independent variables and coefficients

$$f(x_n) = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_j x_{jn} + \cdots + \beta_k x_{kn} \quad (2)$$

Where, Is the x_{jn} vector of the independent variable, and β is the corresponding coefficient.

When there are two or more independent variables in the experimental study, the effect of one of the independent variables on each level of the other is inconsistent. This phenomenon is called the interaction effect. This study takes the first-order interaction effect as the research object, which is limited to two explanatory variables. Therefore, the function It can be expressed by the following expression:

$$f(X_n) = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots + \beta_K X_{Kn} + \beta_{K+1} X_{1n} X_{2n} + \cdots + \beta_M X_{Kn} X_{(K-1)n} \quad (3)$$

Where: K is the number of independent variables; m is the number of variables and interaction effects.

$$M = K(K+1)/2 \quad (4)$$

2.2 Random forest (RF)

RF is one of the most commonly used classifiers proposed by Breiman for training and predicting samples. Based on bootstrap sampling method, RF algorithm can change the training set and establish decision tree set. Because the classification tree is constructed with the guidance of data, and the candidate variable set is a random subset of variables at each split.In this paper, the Gini index of random forest algorithm is used to analyze the influence of different factors on the severity of traffic accidents.[5][6]

The Gini index is calculated as follows:

$$GINI(p) = \sum_{K=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} {p_k}^2 \quad (5)$$

Where: K represents K categories, Represents the sample weight of class K.

So the characteristics The importance on node m, that is, the Gini index changes before and after node m branching are as follows:

$$VIM_{jm}^{(GINI)} = GI_m - GI_l - GI_r \quad (6)$$

Where Gil and Gini index of the two new nodes after branching respectively

If the feature If the nodes in decision tree I are in set M, then the importance of X in the ith tree is

$$VIM_{ij}^{(GINI)} = \sum_{m \in M} VIM_{jm}^{(GINI)} \quad (7)$$

Suppose there are n trees in RF, then

$$VIM_{j}^{(GINI)} = \sum_{i=1}^{n} VIM_{ij}^{(GINI)} \quad (8)$$

Finally, all the obtained importance scores are normalized

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i} \quad (9)$$

The denominator is the sum of all characteristic gains, and the molecule is the Gini index of characteristic J.

2.3 Spatial analysis method

Compared with the traditional statistical (Poisson) model, the spatial location identification of traffic accident

hotspots uses the spatial attributes of accident points. Hot spot analysis results of traditional accident prone section identification will determine that a single intersection or section with high accident rate is a dangerous section, while the accident hot area will identify multiple continuous single road sections as a hot spot area on the basis of considering the spatial autocorrelation of the spatial agglomeration of accident points. The traditional spatial analysis method (hotspot analysis) of traffic accidents may be affected by random factors rather than by road environment. Therefore, in order to explore the main factors of the distribution of traffic accident hot spots, this paper comprehensively considers the following two points: (1) the historical location of traffic accidents; (2) the spatial attributes of the historical locations of traffic accidents.[7]

In this paper, the spatial analysis method based on Arc GIS software is used to evaluate the potential mutual dependence between the attribute values of observation data in a certain analysis range. If the similarity of the observed values of each spatial point becomes more similar with the reduction of spatial distance, it is spatial positive correlation, otherwise it is spatial negative correlation; if there is no obvious relationship between the observed values and spatial relationship, it is spatial uncorrelated

3 Study case

Most of the existing studies show that both macroscopic and microscopic considerations are conducive to the analysis of traffic accidents. Therefore, this study analyzes from two levels.[8]

The map of Beijing is divided into 2023 traffic areas as analysis units. In addition to using ArcGIS software to obtain the basic information of traffic district such as area, edge length and location, this paper also obtains the point of information (POI) data of supermarkets, banks, school supermarkets and office buildings from Google map, and uses mobile phone signaling data to obtain the employment and living conditions of each traffic district. Using the spatial analysis tool of ArcGIS software, the POI data of supermarkets, banks, school supermarkets, office buildings, the number and density of residents and working population in each traffic district are collected and calculated. Figure 1 shows the spatial distribution of the relevant POIs in this paper.

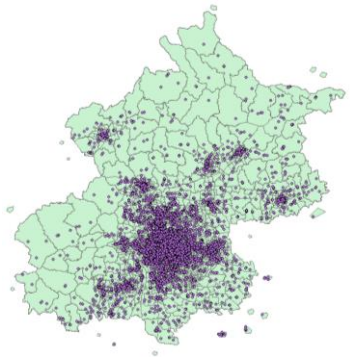

Figure 1. Distribution of Schools.

The macro factors of traffic accidents studied in this paper are as follows:

Table 1. Statistical table of macro factors in study area

Variable	Description	Max	Min	Mean	S.D
Length	The length of TZA(km)	123.22	1.58	9.86	14.42
Ln Work	Work population	9.21	4.61	7.3	0.92
Ln Live	Resident population	9.80	4.80	8.31	0.95
AFC	The dense of subway station	8.34	0.00	0.27	0.77
Bus station	The dense of bus station	306.55	0.00	24.69	35.30
Bank	The dense of bank	81.75	0.00	4.24	8.73
Supermarket	The dense of supermarket	392.55	0.00	8.95	19.04
Restaurant	The dense of restaurant	140.12	0.00	10.65	18.57
Office	The dense of office	68.95	0.00	1.47	4.39
School	The dense of school	51.03	0.00	2.77	4.38
Area	Area of TAZ (km^2)	382.05	0.14	8.61	31.46
Road	The length of road inside TAZs (km)	244.48	0.274	18.326	24659

3.1 Introduction to traffic accident data

In this study, the normal traffic accident data recorded by Beijing police in 2015 were extracted.There are three key components of the collision data: (1) the location during the collision; (2) the collision attributes, including the involved vehicles, pedestrians, casualties, road cross-section, central isolation facilities, road safety attributes, road conditions, road types, road alignment; (3) weather conditions and other environments.

Among the casualties, there is information about the number of casualties and the characteristics of death. In 2015, except for minor accidents, there were 3982 traffic accidents in Beijing. The road type information in traffic accidents is shown in the following table:

3.2 Analysis of accident influencing factors

The graph shows the order of importance of macro factors in Gini coefficient. The top 10 most important factors include the density of restaurants, the density of bus stops, the area and length of TAZ, the density of banks, the

Table 2. Traffic accident data road type statistics

Variable	Description	Percent(%)
Road type	1 Expressway	5.07
	2 Level 1	8.04
	3 Level 2	10.12
	4 Level 3	5.7
	5 Level 4	5.5
	6 Others	1.46
	7 City Expressway	5.83
	8 General City Road	29.03
	9 Units self-built road	0.6
	10 Public parking lots	0.13
	11 Public square	0.03
	12 Other roads	28.5

resident population, the type of roads, and the density of supermarkets. A large number of studies have shown that commercial gatherings affect traffic safety. In terms of all poi densities, restaurants, banks and supermarkets are considered to be important features. At the same time, dense bus stops have become an important macro feature of traffic accidents. In similar studies, there may be more potential risks at bus stops. At present, not all bus stops in Beijing have guardrails to protect waiting passengers, so there is a greater risk of traffic accidents.[9]

Previous studies have discussed a positive correlation between population and traffic accidents. It is confirmed in this paper that more permanent population means more traffic activities. Therefore, in order to improve road safety, we should pay special attention to the area with relatively dense population, which may become the hot spot of traffic accidents.[10]

In addition, school density is also an important factor affecting traffic accidents, because schools can attract a lot of traffic, so the relatively large traffic attraction will be accompanied by traffic accidents in the suburbs. Therefore, it is necessary to study the school density as an important factor in the identification of traffic accident hot spots.

In addition, road type is very important in micro factors, so binary logistic model is used to further analyze road type, and different symbols are used to indicate the degree of PR (> | z|): "* *" represents 0.001-0.01, "*" represents 0.01-0.05, "." represents 0.05-0.1, and "*" represents 0.1-1, as shown in Table 2.

Table 3 shows class IV highways and general urban roads and serious pedestrian accidents. In addition, the first-class highway and self -built road are also related to the accident.

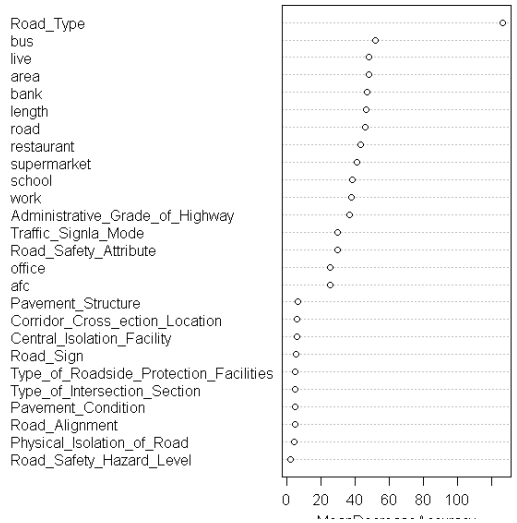

Figure 2. Results of Gini coefficient of accident influencing factors

Table 3. Statistics of influencing factors of road type in binary logic model

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Sign
First class road	-2.01048	0.79535	-2.528	0.011479	*
Second class road	-1.48124	0.76465	-1.937	0.052725	.
Third class road	-2.1823	0.79753	-2.736	0.006213	* *
Fourth class road	-4.07199	0.95797	-4.251	2.13E-05	* * *
General city road	-2.57305	0.7248	-3.55	0.000385	* * *
Self-built road	-2.73951	1.2744	-2.15	0.031584	*
Other roads	-2.25638	0.73417	-3.073	0.002116	* *

3.3 Identification of traffic accident hot area

The previous section has analyzed and introduced the influencing factors of traffic accidents from macro and micro aspects in detail, carried out correlation analysis on many related factors, and determined the spatial characteristics of dangerous road sections. This section mainly introduces the process of hot zone identification by using the influencing factors of traffic accidents, so as to provide strong scientific basis for improving urban traffic safety.

Figure 3. Traffic accident distribution map of the study area

This paper selects the inner area of the Fourth Ring Road as a case of traffic accident hot area identification for further analysis. The traffic accident hot zone is composed of a series of continuous road sections with a high number of traffic accidents. According to the above analysis of traffic accident influencing factors, the historical location of traffic accident, road type, supermarket density, school density and station density are selected as important indicators to identify the hot area. The specific flow chart of the identification process is shown in Figure 4.Firstly, the core density tool and linear density tool of ArcGIS software are used to identify the density of POI points, and then the density of each POI point is re sorted by the re classification tool in ArcGIS software according to the equal interval. The new classification categories are 10, and the higher the density, the higher the score. Finally, weighted stack is carried out. According to the order of Gini coefficient, the weight is set as shown in Table 4.

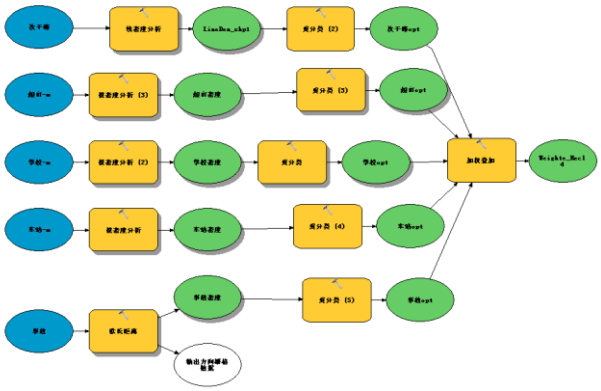

Figure 4. Flow chart of Arc GIS identification of accident hot zones

The results of hot zone identification are shown in Fig. 5.It can be seen from Figure 5 that the hot spots of traffic accidents are scattered, relatively concentrated in the area from the second ring road to the Third Ring Road in the urban area of Beijing. Figure 6 shows the nuclear density map of traffic accidents. Traffic accidents are distributed intensively, and key traffic accident dangerous areas cannot be identified.

Table 4. The fixed number and carrying capacity of passengers

Serial number	factor	weight
1	Historical accident location	30
2	Road grade	30
3	Business density	10
4	School density	10
5	Density of bus stops	20
the sum		100

Compared with figure 5 and Figure 6, it can be seen that the traffic accident hot area in Figure 5 overlaps with the area where the traffic accident is concentrated as shown in Figure 6, but there is no great degree of risk differentiation. At the same time, if we don't consider the influence factors of spatial attributes among the traffic accident points, almost all the traffic accident points with high accident frequency will be considered as the accident hot area, which indicates that even the area with relatively concentrated traffic accident points in the historical data is not necessarily the potential traffic accident hot area. On the contrary, many areas with no high incidence of traffic accidents in history have the possibility of becoming high-risk areas of traffic accidents. Further inspection is needed to eliminate potential safety hazards.

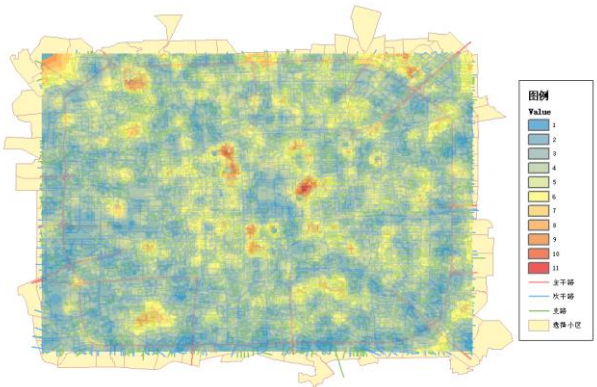

Figure 5. Identification results of traffic accident hot zones

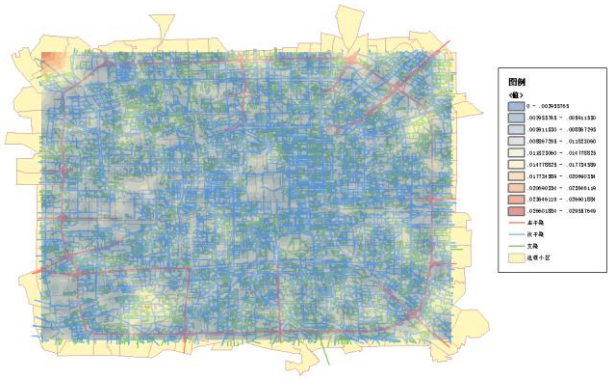

Figure 6. Traffic accident distribution map

4 Conclusion

1) In this paper, considering the attributes of the accident itself and the spatial attributes, the identification method of traffic accident hot area is proposed, which is of great significance to improve road traffic safety and formulate targeted improvement measures and suggestions.

2) In this paper, random forest and binary logistic model are used to identify the macro and micro influencing factors of traffic accidents. The influence of road infrastructure, road spatial environment and socio-economic environment on the distribution of accident hot spots is comprehensively considered. Spatial statistics and mathematical statistics are combined to comprehensively consider the attributes and spatial attributes of accidents.At the same time, the priority level of hot area was set up, which was divided into 10 grades to rank the severity and risk.In general, in order to ensure a higher efficiency of investment improvement, priority can be given to the adjustment and optimization of the most dangerous accident hot area.

3) In this paper, there are still some deficiencies and areas that can be further improved in the exploration of traffic accident hot area identification: we can further explore the hierarchy of the weight setting of each index in the weighted analysis, and establish a scientific index system to identify the accident hot area.In addition, more factors can be considered to further explore the distribution of traffic accident hot spots.

References

1 Chen Yuefei. Development of identification and analysis system for dangerous highway sections based on ArcGIS [D]. 2015
2 Shen feimin. Road traffic safety. Beijing Machine Press, 2007:219-25
3 Wu xiuxu et al. Application and practice of ArcGis9 geographic information system [M]. Beijing: Tsinghua University Press, May 6, 2007
4 Xiong Li. Research on identification of traffic accident hot area and analysis method of hot area cause based on ArcGIS [D]
5 Besharati, M. M., Tavakoli Kashani, A., Li, Z., Washington, S., & Prato, C. G. (2020). A bivariate random effects spatial model of traffic fatalities and injuries across Provinces of Iran. Accident Analysis & Prevention, 136, 105394. doi:10.1016/j.aap.2019.105394
6 Ramón Díaz-Uriarte, Sara Alvarez de Andrés. Gene selection and classification of microarray data using random forest[J]. Bmc Bioinformatics, 2006, 7(1):3-0.
7 MOONS E. etal Identifying hazardous road locations: Hot spots versus hot zones[J]. Trans. Computational Science, 2009, 6: 288–300.
8 Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., & Abdel-Aty, M. (2016). Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of Transport Geography, 54, 248–256.doi:10.1016/j.jtrangeo.2016.06.012
9 Noland, R.B., Quddus, M.A., 2005. Congestion and safety: a spatial analysis of London.Transport. Res. A 39, 737–754
10 Yang K , Yu R , Wang X , et al. How to determine an optimal threshold to classify real-time crash-prone traffic conditions?[J]. Accident Analysis & Prevention, 2018, 117(AUG.):250-261.