# The traffic accident hotspot prediction: Based on the logistic regression method

Tao Lu
GNSS Research Center
Wuhan University
Wuhan, 430079, China

Yan Lixin
Intelligent Transport Systems Research Center
Wuhan University of Technology
Wuhan, 430063, China

Zhu Dunyao
Intelligent Transport Systems Research Center, Wuhan University of Technology
GNSS Research Center, Wuhan University
Wuhan, 430063, China

Zhang Pan
GNSS Research Center
Wuhan University
Wuhan, 430079, China

**Abstract—Traffic accident has posed a threat to the safety of human life. Accident often occurs in some certain areas. In order to achieve the target of the identification and judgment of the accidents hotspots during driving, and improve the driving safety of vehicle and the traffic efficiency through early warning, a research was done based on the recent collected 400 sets of accidents data of 10 major roads in Beijing city. Through the statistics of the typical factors, and the Logistic regression analysis, the relationships between the traffic accident and the road type, the vehicle type, the driver state, the weather, the date etc., were studied. Finally, the prediction model of accident hotspot was established. The results show that the location of car in road transects, the road safety grade, the road surface condition, the visual condition, the vehicle condition and the driver state are the most significant factors which may lead to traffic accident. Meanwhile, the prediction model established in this paper was validated to be capable of predicting the occurrence of accident, and the prediction accuracy is approximate 86.67%. The study provides not only a theoretical basis for vehicle safety assistance driving, but also the guidance for collision avoidance and optimization for path planning of intelligent vehicle.**

***Keywords—traffic accident hotspot; logistic regression analysis; accident prediction; safety assistance driving***

## I. INTRODUCTION

Currently, over 130 million people die worldwide each year in road accidents caused by cars; meanwhile road accident has become a major killer and posed a threat to human life[1]. The occurrence of traffic accidents may attribute to several factors, such as drivers, vehicles, roads, environment, and so on. Therefore, if one can find out the relationship between traffic accidents and various factors, the probability of accidents can be reduced effectively through early warning before accident, and it will be very helpful to traffic safety.

With the development of intelligent transportation systems and active safety technologies, the prediction method of traffic accident has become a hot topic. At present, there are several accident prediction methods such as regression analysis, gray theory method, time series method, neural network etc. Fang and Shen analyzed the linear relationship between average annual accident death number and GDP, motor vehicle population, highway mileage, and population and by using multiple regressions[2]. Li, Zhang and Jiang introduced markov chains forecasting method into grey forecasting method, and then built the grey-markov model for forecasting the average annual number of accidental death [3]. Liu, Ren and Duan established the time series model of average annual accidental death number[4]. Li and Shao, using BP neural network, analyzed the relationship between the factors, including driver population, motor vehicle population, trunk mileage, weather etc., and average annual accidental death- and injured- number [5].

But all the above researches only macroscopically analyze the death or injured number of average annual accidents. And the established models can't be used in microscopic traffic accident prediction. Qin, Liu and Zhang built the microscopic forecasting model using Bayesian network, with weather, time, traffic flows and vehicle speed as the independent variables[6]. But in this model, although vehicle speed is related to motor vehicle and the rest three parameters reflects the state of environment, it ignored that driver and road also affect traffic accident significantly. Luo, Li, Hu, Liu and Zheng through factor analysis and logistic regression, built a traffic accident forecasting model considering visibility, humidity, precipitation and other meteorological factors as independent variables[7]. However, the model took only meteorological factors into account, and ignored the influence of drivers, roads and environment on traffic accident. Guo and Fang identified the factors associate with individual driver risk, and then built a high- and moderate-risk drivers prediction model using principal factor analysis and logistic regression with demographic, personality and characteristic data[8]. Although the main reason of traffic accident is the driver in most cases, the neglect of other factors in accident, would not meet the

objective facts, and would pose an adverse effect on accident prevention [9].

Therefore, this paper investigated a number of traffic accidents data of Beijing's 10 main roads. And select 6 independent variables from people-vehicle-environment diverse perspective. By using logistic regression, the relationship between accident and several factors was analyzed, and then a traffic accident hotspot prediction model was established which would provide the theoretical basis of traffic accident warning, intelligent car collision avoidance and vehicle active safety technology.

## II. STATISTIC METHOD

Assuming that there are n factors affecting whether accidents occur, denoted by: $x_1$ , $x_2$ ,…, $x_n$ . The logistic model is:

$$\log it(y) = \ln(\frac{p}{1-p}) = a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n \quad (1)$$

Where $y$= (0, 1) is the judgment of traffic accident hotspot; $p$ is the probability of traffic accident's occurrence; $x_i$ ($i$=1,2,…,n) is the factor associated with traffic accident; $a_0$ is the constant, and $a_i$ ($i$=1,2,…,n) are regression coefficients. The model could equivalently be expressed as:

$$p = \frac{e^{(a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n)}}{1 + e^{(a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n)}} \quad (2)$$

## III. THE IMPACT FACTORS OF ACCIDENT HOTSPOT

### A. The selection of independent variables

The traffic accident data investigated in this paper includes detailed traffic accident statistical information, such as: location of the accident, road grade, time, accident cause, light condition, road surface condition, number of fatalities, and so on.

A dynamic equilibrium driving system is made up by the driver, the car, the road and the environment. Any element's failure of this system will lead to traffic accident, and the incentives could be subjective or objective or both. Generally, the vehicle driver is regarded as subjective factor while the road, the vehicle, the environment are objective factors. Even though most traffic accidents are caused by the drivers, however, the effect of road, vehicle, and surrounding can't be ignored. Therefore, when selecting the factors of traffic accident hotspot, not only the subjective factors but also the objective factors should be taken into consideration. In this study, 6 independent variables are chosen form various aspects, as shown in table 1.

TABLE I. INDEPENDENT VARIABLES SELECTED

| Influencing factors | Independent variables selected |
|---|---|
| Road | Location of car in road transects |
| | Road safety grade |
| | Road surface condition |
| Environment | Visual condition |
| Vehicle | Vehicle condition |
| Driver | Driver state |

### B. Analysis of the arguments

All the 6 variables selected are categorical variables, and they will be generalized and abstracted to be 6 dichotomous variables via the following analyses.

*1)* Location of car in road transects. Sections perpendicular to the centerline of the road are known as road transects. In this study's accidents data, the locations in road transect consist of motor lane, bicycle lane, pavement, crosswalk etc. As shown in table 2, all those locations would be classified as conflict area or non-conflict area.

TABLE II. THE GENERALIZATION OF THE LOCATIONS IN ROAD TRANSECTS

| The original variable value | The new variable value |
|---|---|
| Motor lanes | non-conflict area |
| Mixed lanes | |
| bicycle lanes | conflict area |
| pavement | |
| crosswalk | |

*2)* Road safety grade. In China, roads are classified as highways and urban roads. Highways consist of expressway, first-class highway, second-class highway, third-class highway, fourth-class highway and countryside road. Urban roads consist of urban expressway, main road, secondary road, and minor road. The roads' width and lane number vary with the road level. The higher level of the road, the wider the road, with more lanes. So its traffic capacity is stronger and it will be more safe when driving on. The value of influence coefficient on lanes number safety of the 4-lanes road is 1, the larger the coefficient, the greater effect on safe driving[10].Table 3 shows that roads with 4 or more lanes will be deemed as high safety grade roads, and the rest are low safety grade roads.

TABLE III. THE GENERALIZATION OF ROAD GRADE

| The original variable value | The new variable value |
|---|---|
| expressway | high safety grade roads |
| first-class highway | |
| second-class highway | |
| urban expressway | |
| main road | |
| secondary road | |
| third-class highway | low safety grade roads |
| fourth-class highway | |
| countryside road | |
| minor road | |

*3)* Road surface condition. In the statistical data of traffic accident, there are totally 11 types of road surface conditions, such as: moist, flat, muddy, potholed and so on. Poor road surface conditions will adversely affect safe driving. In the 11 kinds of surface conditions, only flat road surface condition is good and the rest are poor.

*4)* Visual condition. Good visual conditions are particularly important for safe driving. In poor visual conditions, the driver's judgment will be severely affected to cause accident. There are 3 visual conditions: daytime, night with street light and night without street light. Obviously, daytime and night with street light are in good visual conditions while night without street light is in poor visual condition.

*5)* Vehicle condition. Vehicle conditions have a direct influence on driving safety. Poor vehicle conditions may lead to serious traffic accidents. There are many vehicle conditions, for example: brake failure, steering failure, light failure and tire's explosion etc. A vehicle with any failure would be defined as poor vehicle condition, otherwise as good vehicle condition.

*6)* Driver state. In all influence factors, particularly, the driver's bad state, which includes: drunken driving, driver fatigue, traffic offence etc., is the major factor that cause accidents directly.

Through the above analyses, the 6 categorical variables are generalized to be 6 dichotomous variables and encoded, as shown in table 4.

TABLE IV. DESCRIPTION OF INDEPENDENT VARIABLES

| Independent variables | Description | Code |
|---|---|---|
| $x_1$ | Location of car in road transects | 0=non-conflict area 1=conflict area |
| $x_2$ | Road safety grade | 0= high safety grade road 1= low safety grade road |
| $x_3$ | Road surface condition | 0=poor 1=good |
| $x_4$ | Visual condition | 0=poor 1=good |
| $x_5$ | Vehicle condition | 0=poor 1=good |
| $x_6$ | Driver state | 0=poor 1=good |

### C. *The definition and selection of dependent variable*

In this paper, to establish the prediction model of traffic accident hotspot, we define $y$ as the dependent variable, which represents whether an area is the traffic accident hotspot, if it is, $y$=0, otherwise, $y$=0.

## IV. RESULTS

### A. *The establishment of traffic accident hotspot prediction model*

In this paper, 170 sets of Beijing's traffic accident data from 2004 to 2007 were selected to study, and then the traffic accident hotspot prediction model was developed.

All the 6 factors were used in logistic regression, and the outputs of the model are summarized in table 5, where B is the parameter estimates of the regression equation; S.E. is the standard deviation of the regression coefficients reflecting the data's discrete degree [11]; Wald is the wald test value which can be used to test the significance of regression coefficients; the greater value, the greater role of the corresponding independent variable; df means the arguments' degree of freedom, and Sig reflects the significance level. From Table5, it can be found that all the 6 independent variables have significant effects on the dependent variable, so they should be used in the logistic model's building.

TABLE V. VARIABLES IN THE LOGISTIC EQUATION

| Independent variables | B | S.E. | Wald | df | Sig. |
|---|---|---|---|---|---|
| $x_1$ | 1.424 | 0.451 | 9.964 | 1 | 0.002 |
| $x_2$ | 1.507 | 0.529 | 8.120 | 1 | 0.004 |
| $x_3$ | 2.043 | 0.591 | 11.948 | 1 | 0.001 |
| $x_4$ | -2.542 | 0.747 | 11.585 | 1 | 0.001 |
| $x_5$ | 2.004 | 0.942 | 4.528 | 1 | 0.033 |
| $x_6$ | -1.378 | 0.438 | 4.528 | 1 | 0.002 |
| *Intercept* | -1.547 | 1.193 | 1.681 | 1 | 0.195 |

The traffic accident hotspot prediction odds ratio model is as follow:

$$\mathrm{logit}(y)=\ln\left(\frac{p}{1-p}\right)=-1.547+1.424x_1+1.507x_2+2.043x_3-2.542x_4+2.004x_5-1.378x_6 \tag{3}$$

And the probability model is:

$$p=\frac{e^{(-1.547+1.424x_1+1.507x_2+2.043x_3-2.542x_4+2.004x_5-1.378x_6)}}{1+e^{(-1.547+1.424x_1+1.507x_2+2.043x_3-2.542x_4+2.004x_5-1.378x_6)}} \tag{4}$$

### B. *Validation of the model*

60 sets of Beijing's traffic accidents data in 2004~2007 were selected for model validation, in which 8 sets of the data were misjudged to be non-accident hotspot and 52 sets of the data were judged to be accident hotspot correctly with 86.67% correct rate, as shown in fig.1.

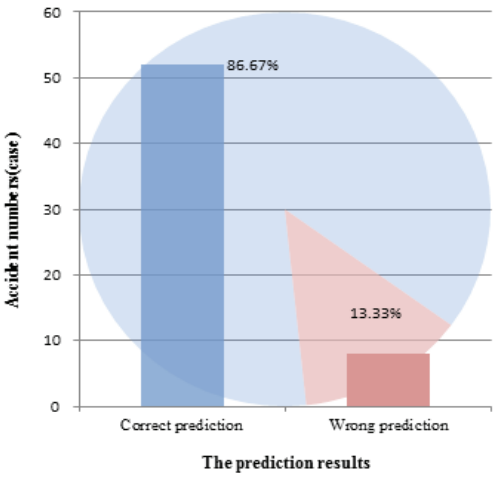


Fig. 1. The results of model validation

The traffic accident is affected by many factors, i.e., weather, traffic flows, road alignment etc. The occurrence of traffic accident is complex and random. Even though the model built in this paper could not express the essential characteristics of traffic accident, it can provide the relatively high accuracy in accident prediction.

## V. CONCLUSIONS

In this paper, 6 factors which significantly associate with traffic accident were developed from multiple aspects, which are the driver, the road, the car and the environment. On this basis, a logistic traffic accident hotspot prediction model was established with the location of car in road transects, the road safety grade, the road surface condition, the visual condition, the vehicle condition, and the driver state as the independent variables and traffic accident hotspot as the dependent variable. The results show that this model could provide good predictions against traffic accident with 86.67% correct rate. It should be noted that due to the constraints of data and research condition, there are still some factors, such as weather, traffic flows etc., not used in the model and they should be taken into account in future study. The results of this study can be used in vehicle safety assistance driving and provide early warnings and proposals for safe driving.

## REFERENCES

[1] Yang X.X., Zhang X., Chang G.X., Lou Y.Y., (2008). "Development Status and Trends of Automobile Active Safety Technology." Journal of Chongqing Institute of Technology (Natural Science)., 22(4), 15-17.

[2] Fang Y.R., and Shen F.M., (2012). "Development trend analysis and prediction of traffic accident." Journal of Safety Science and Technology. 8(3), 141-146.

[3] Li X.Y., Zhang N., and Jiang G.F., (2003). "Grey-markov Model for Forecasting Road Accidents." Journal of highway and transportation research and development, 20(4), 98-100.

[4] Liu X.M., Ren F.T., Duan H.L., (1995). "Road accident forecasting in time series" China journal of highway and transport. 8(S1), 125-130.

[5] Li J. and Shao C.F., (2006). "The traffic accidents prediction model based on BP neural network." Computer and Communications. 24(2), 34-37.

[6] Qin X.H., Liu L., and Zhang Y., (2005). "A traffic accident prediction method based on Bayesian network model." Computer Simulation. 22(11), 230-232.

[7] Luo H., Li L.X., Hu S., Liu Y. and Zheng L., (2007). "The relationship between road traffic crashes and meteorological condition with construction of its road weather warning model." Journal of Applied Meteorological Science. 18(3), 350-357.

[8] Guo F., and Fang Y., (2013). "Individual driver risk assessment using naturalistic driving data." Accident Analysis & Prevention, 61, 3-9.

[9] Ding Z.L., and Zheng Y., (2010). "Revelation of traffic safety on in-depth investigation for road accident." China Public Security. Academy Edition. (1), 86-90.

[10] Pei Y.L., and Ma J., (2003). "Research on countermeasures for road condition causes of traffic accidents." China Journal of Highway and Transport. 16(4), 77-82.

[11] Ning M.D., (1994). "Integral realization and application on standard error." Journal of national university of defense technology. 16(3), 131-134.