Research on Short-term Prediction Model of Freeway Operation Situation

Xiaodan Zhang[1]

[1] *Highway Traffic Safety Engineering Research Center, Research Institute of Highway Ministry of Transport, Beijing, China*

Abstract. Based on the traffic flow data and accident data of Beijing-Tianjin-Tanggu freeway, the security situation short-term prediction model was established in the paper. Firstly, we established the risk prediction database, and developed the pre-analysis software of basic data; secondly, the traffic flow data between 10 to 15 minutes prior to the time of accident were aggregated at 5-minute level, and the volume, speed, occupancy as well as their statistical parameters were selected; finally, based on the correlation analysis results of parameters, the multi-parameters Logistic regression model was established. The results indicate, the change of traffic flow parameters and their statistics can effectively predict the possibility of accident, in which the average value of speed of small car, the standard deviation of volume of large car and the average value of volume difference between large car and small car at 5-minute level have a significant impact on the risk of accident.

1 Introduction

Intelligent transportation system (ITS) is widely used in the global scope, which makes the traffic managers have large real-time traffic situation data. Many researchers and practitioners have been fully aware that all the advantages of ITS will not be recognized without realizing the ability of the traffic flow short-term prediction[1] (Brian L. Smith, 2002). The traffic flow prediction model can provide such a kind of ability, and can provide forward-looking traffic management as well as comprehensive travel information service.

At present, the object of traffic forecast is mainly urban road, and it focuses on the unblocked reliability and efficiency[2-7], including the assessment of traffic operation situation, short-term traffic forecast, road traffic condition judgment and so on, but the researches on the traffic flow operation security situation prediction are rare.

In China, due to the lack of detailed accident data and microscopic traffic flow data as support, resulting in a serious shortage in the real-time traffic flow security analysis theory, leading to the current freeway security management lag behind the real-time traffic situation prediction in our country. The United States and Canada began to carry out researches on the traffic accident detection algorithms and the traffic flow harbinger characteristics before the traffic accident from 1990s[8-15]. Among them, Chris adopted the speed differences between upstream and downstream, and the variances of the cross-section speed as the characterization factors of the traffic flow real-time risk discrimination[9], the results of which was referenced by Kansas state highway agency of USA in 2006. However, the main shortcomings were unknown of the risk reason and subjectivity of the risk rank assessment, and its scientificity needs to be further examined.

So, the prediction method of freeway operation situation based on short-term traffic flow multi-parameters regression is researched in this paper, thus achieving the short-term prediction of the traffic flow operation situation, and the results are helpful to reduce the risk of accident, decrease traffic accidents, and improve the operation security of freeway.

2 Data processing

Beijing-Tianjin-Tanggu freeway is about 142.69 kilometers, has 28 microwave detectors, and the average spacing of detectors is 5 kilometers, in which, the coils of Beijing section is more intensive, but some of the data is missing. All the traffic flow data as well as the traffic accidents data in Beijing-Tianjin-Tanggu freeway are extracted in this paper, in which, the traffic accidents include time, location, type, reason, and et. al., and the traffic flow data are speed, volume and occupancy of divided-lane at 1-minute level.

First, we select the nearest detector data within 2 kilometers prior to or back to the location of accident, in order to screen data for first round, in which, the choice of accidents should avoid these caused by external factors such as weather, linear conditions, drivers and vehicles as much as possible, only in this way we can accurately excavate the rule of the traffic flow fluctuation affecting on the accidents, and in consideration that the relationship between the causes and effects of single-vehicle accidents and the traffic flow may not be strong under high service level, so the select of accident samples is more emphasis on the multi-vehicle accidents in large traffic volume (under C-classical service level); and then, we screen data for second round: check the quality of the traffic flow data, including deleting and processing the singular values such as speed with 0 km/h, diagnose the outliers by the data spatio-temporal graph method and the statistical method, and correct the abnormal data values by the simple difference method and the filtering method, in order to improve the accuracy and reliability of the analytical results, thus selecting the speed, volume and occupancy of detectors with better data quality to pair with accidents, and extract the traffic flow data of the control groups by 1:4 ratio.

The control group data meet the following requirements: the date is different with the corresponding accident; the time, week, and location are the same as the corresponding accident; the control group has no accidents at this location at the same day. Then, we select the control groups with better data quality according to the same method mentioned above, so as to establish the database required in this paper.

Because the location of traffic accident recorded by the police department is a cross-sectional stake number, therefore, it is necessary to aggregate the divided-lane traffic flow data into a cross-sectional traffic flow data, thus taking the cross-sectional data as the foundation of the research, so, the divided-lane data are transformed into the cross-sectional data by weighting method: $q_s = \sum_{i=1}^{n} q_i$, $k_s = \frac{\sum_{i=1}^{n} q_i \cdot k_i}{\sum_{i=1}^{n} q_i}$, $v_s = \frac{\sum_{i=1}^{n} q_i \cdot v_i}{\sum_{i=1}^{n} q_i}$, in which, q_i , k_i , v_i are respectively volume, occupancy, speed of divided-lane, n is the number of lanes.

3 Mathematical model

3.1 Logistic regression model

Binary Logistic regression model is commonly used to quantitatively analyze the impact of explanatory variables on binary dependent variables, also can be used to estimate the occurrence probability of a category of the dependent variables, and from the traffic flow operation results, the dependent variables just can be divided into two categories: accidents and non-accidents. The probability of the accident corresponding to one sample data is:

$$P(x_i) = \frac{1}{1+e^{-x_i^{'}\beta}}, \quad i=1,2,\cdots,n \tag{1}$$

The linear expression after logit transform is:

$$\ln\frac{P(x_i)}{1-P(x_i)} = x_i^{'}\beta, \quad i=1,2,\cdots,n \tag{2}$$

Where, $P(x_i)$ represents the probability of traffic accident; $x_i^{'}\beta$ represents the linear combination of explanatory variables: $x_i^{'}\beta = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$, x_{ki} is the value of variable k in i -th sample; β_0 is the regression intercept; $\beta_1, \beta_2, \cdots, \beta_k$ are the regression coefficients of explanatory variable x_{ki} ; $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ can be calculated by the maximum likelihood estimation method:

$$\ln L(\beta, x_i) = \sum_{i=1}^{n} [\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} - \ln(1+e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}})] \,.$$

3.2 Logistic model testing

In Logistic regression, likelihood ratio test, Akaike information criterion (Akaike Information Criterion, AIC) and Schwarz criterion can be used to reflect the goodness of the model fitting.

We adopts AIC to reflect the fitting effect of the final model in this paper, that is $AIC = -2LL + 2(K+S)$, in which, K is the number of independent variables of the model, S is the total number of response variable categories minus 1, the range of $-2LL$ is 0 to infinity, which is the smaller the better. Under the same conditions, the smaller value of AIC indicates the better model fitting.

The prediction accuracy for classification is usually used to reflect the prediction accuracy of the model. Using Logistic model to predict the classifications needs specifying a threshold of probability, that is, when the probability calculated by Logistic model is greater than a specified threshold, it is discriminated as the traffic accident, and when the probability is less than a specified threshold, it is discriminated as the security state with non-accident. The threshold value decides the forecast accuracy of each category and the total samples, the current researches commonly use the proportion of a category in the whole samples as the threshold value of this category prediction. Because of researching the prediction method of freeway traffic operation situation short-term in this paper, the proportion of accidents in the whole samples is adopted as the threshold value.

4 Research results and analysis

4.1 Data preparation Logistic regression model

In order to predict the traffic accident in advance, the calibrated traffic flow data within 10~15 minutes prior to the time of the accident is extracted in this paper, meanwhile, the traffic flow data of the control group in the corresponding period is extracted for each accident. Through screening of data quality, we ultimately retain 33 groups of accident samples and 132 groups of non-accident samples as the control groups for modeling research, and divide them into 2 categories: accident and non-accident, that is the value of dependent variable is 1 indicating accident and 0 meaning non-accident.

Firstly, the original traffic flow data of divided-lane are aggregated into a cross-sectional traffic flow data, secondly, in order to avoid the data noise due to the short acquisition interval, the data are converged with 5-minute level in order to get averages and standard deviations.

4.2 Model Parameters Selection

We extract 14 statistical parameters as the chosen model parameters, as shown in Tab. 1:

Table 1. Model parameters

Variable Symbol	Symbolic Interpretation
lvavg	the average value of speed of large car at 5-minute level
svavg	the average value of speed of small car at 5-minute level
lvsd	the standard deviation of speed of large car at 5-minute level
svsd	the standard deviation of speed of small car at 5-minute level
gavg	the average value of occupancy at 5-minute level
gsd	the standard deviation of occupancy at 5-minute level
lqavg	the average value of volume of large car at 5-minute level
sqavg	the average value of volume of small car at 5-minute level
lqsd	the standard deviation of volume of large car at 5-minute level
sqsd	the standard deviation of volume of small car at 5-minute level
vdavg	the average value of speed difference between large car and small car at 5-minute level
vdsd	the standard deviation of speed difference between large car and small car at 5-minute level
qdavg	the average value of volume difference between large car and small car at 5-minute level
qdsd	the standard deviation of volume difference between large car and small car at 5-minute level

4.3 Modelling steps

The binary Logistic model of the freeway traffic security operation situation deduction based on short-term traffic flow multi-parameters is established by using statistical analysis software with R programming language in this paper, the steps are as follows:

(1) We use the correlation analysis method to examine the correlation between variables, which makes the highly correlated variables be not into the Logistic model;

(2) We select the reasonable explanatory variables for modeling according to the backward selection method of the Logistic regression, the basic steps are as follows:

a) First, all the variables are contained in the model;

b) Then, calculate the z test values of all variables, and get the corresponding P values;

c) Last, find the largest P value, if the P value is greater than the significance level α_{out} , this variable is eliminated;

d) Go back to step b) for the next round of elimination.

Among them, the significance level of the reserved values is set as: including variables is $P \le 0.05$, excluding variables is $P > 0.1$.

4.4 Logistic model regression and analysis

According to the above modeling steps (1), the results are shown in Tab. 2.

Generally, if the correlation coefficient between two parameters is > 0.6 or < -0.6, they are strongly related to each other, and they cannot enter the model at the same time. It can be seen that:

lvsd has a strong correlation with *sqsd*, *vdsd*, *vdavg*, and *lqavg*;

gavg has a strong correlation with *gsd*, *sqavg*, and *qdavg*;

gsd has a strong correlation with *gavg*;

lqavg has a strong correlation with *lvsd*;

sqavg has a strong correlation with *qdavg*, *qdsd*, *lqsd*, and *gavg*;

lqsd has a strong correlation with *sqavg*, and *qdsd*;

sqsd has a strong correlation with *vdsd*, *lvsd*, and *vdavg*;

vdavg has a strong correlation with *sqsd*, *vdsd*, and *lvsd*;

vdsd has a strong correlation with *sqsd*, *lvsd*, and *vdavg*;

qdavg has a strong correlation with *sqavg*, *qdsd*, and *gavg*;

qdsd has a strong correlation with *sqavg*, *qdavg*, and *lqsd*.

By comparing, we select *lvavg*, *svavg*, *lvsd*, *svsd*, *gsd*, *lqsd*, and *qdavg* as the variables of the model, and according to the above modeling steps (2), the results are as follows in Tab. III:

It can be seen from the z value, the significance levels of parameters are all $P < 0.05$, indicating that *svavg*, *lqsd* and *qdavg* have significant effect on the traffic accident risk of the freeway.

Odds ratio of the traffic flow parameters can be used to quantify the impacts of different traffic flow parameters on the risk of accident. The odds of an event is defined as the ratio of the probability with occurrence and the probability without occurrence, therefore, for the Logistic regression model:

$$\log[\frac{\Pr ob(event)}{\Pr ob(nonevent)}] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{3}$$

$$Odds = \frac{\Pr ob(event)}{\Pr ob(nonevent)} = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k} \tag{4}$$

It can be seen that when the i-th independent variable changes a unit, the change of odds is $Exp(\beta_i)$, that is:

$$Odds\ ratio = \frac{Odds(X_i + 1)}{Odds(X_i)} = Exp(\beta_i) \tag{5}$$

If the coefficient of an independent variable is positive, it means that odds will increase, and this value will be greater than 1; if the coefficient of an independent variable is negative, it means that odds will decrease, and this value will be less than 1; when the coefficient of an independent variable is 0, this value is equal to 1. The change percentage of odds ratio is $Exp(\beta_i) - 1$.

From Tab. 3, the odds ratios of *svavg*, *lqsd* and *qdavg* are respectively 0.96659, 1.49283 and 0.84473, indicating that when *lqsd* changes a unit, the risk of traffic accident will increase 49.3%, however when *svavg* and *qdavg* change a unit, the risk of traffic accident will decrease 3.3% and 15.5% respectively.

According to the analysis, when *lqsd* increases, it indicates that the difference between the volume of large car and the average value of volume of large car increases, that is the distribution of volume of large car is more discrete, which easily causes the traffic flow be not stable: congestion state a while, or free flow state a while,

Table 2. Correlation coefficients of variables

	lvavg	*svavg*	*lvsd*	*svsd*	*gavg*	*gsd*	*lqavg*	*sqavg*	*lqsd*	*sqsd*	*vdavg*	*vdsd*	*qdavg*	*qdsd*
lvavg	1	0.4937	-0.5472	-0.0728	-0.4874	-0.4464	0.3994	-0.1670	-0.0490	-0.4065	-0.3287	-0.4065	-0.1666	-0.2354
svavg	0.4937	1	-0.0094	-0.4246	-0.5947	-0.4970	0.0672	-0.1815	-0.0063	0.0651	0.1899	0.0651	-0.2569	-0.1977
lvsd	-0.5472	-0.0094	1	0.1581	-0.1039	-0.0828	-0.6033	-0.0393	-0.0303	0.8476	0.6059	0.8476	-0.0575	0.0688
svsd	-0.0728	-0.4246	0.1581	1	-0.2486	-0.1071	-0.2588	-0.4052	-0.4437	0.3720	0.4672	0.3720	-0.2678	-0.3330
gavg	-0.4874	-0.5947	-0.1039	-0.2486	1	0.6510	0.0584	0.6334	0.4389	-0.2841	-0.4409	-0.2841	0.6128	0.5728
gsd	-0.4464	-0.4970	-0.0828	-0.1071	0.6510	1	-0.0237	0.1312	0.0928	-0.1915	-0.3110	-0.1915	0.1591	0.1178
lqavg	0.3994	0.0672	-0.6033	-0.2588	0.0584	-0.0237	1	0.0833	0.1247	-0.5648	-0.5534	-0.5648	0.0930	-0.0139
sqavg	-0.1670	-0.1815	-0.0393	-0.40515	0.6334	0.1312	0.0833	1	0.6650	-0.2170	-0.3912	-0.2170	0.9164	0.8637
lqsd	-0.0490	-0.0063	-0.0303	-0.4437	0.4389	0.0928	0.1247	0.6650	1	-0.1653	-0.3398	-0.1653	0.5984	0.6192
sqsd	-0.4065	0.0651	0.8476	0.3720	-0.2841	-0.1915	-0.5648	-0.2170	-0.1653	1	0.7182	1	-0.2022	-0.0787
vdavg	-0.3287	0.1899	0.6059	0.4672	-0.4409	-0.3110	-0.5534	-0.3912	-0.3398	0.7182	1	0.7182	-0.3562	-0.2534
vdsd	-0.4065	0.0651	0.8476	0.3720	-0.2841	-0.1915	-0.5648	-0.2170	-0.1653	1	0.7182	1	-0.2022	-0.0787
qdavg	-0.1666	-0.2569	-0.0575	-0.2678	0.6128	0.1591	0.0930	0.9164	0.5984	-0.2022	-0.3562	-0.2022	1	0.6598
qdsd	-0.2354	-0.1977	0.0688	-0.3330	0.5728	0.1178	-0.0139	0.8637	0.6192	-0.0787	-0.2534	-0.0787	0.6598	1

Table 3. Regression analysis results

Variable	**Coefficient** (β_i)	**Standard error**	***z*** **value**	**Significance level** ($\Pr(>\lvert z \rvert)$)	$Exp(\beta_i)$
Constant	0.48457	0.78750	0.615	0.04383	1.62348
svavg	-0.03398	0.01080	-3.145	0.00166	0.96659
lqsd	0.40067	0.15989	2.506	0.01221	1.49283
qdavg	-0.16874	0.08121	-2.078	0.03772	0.84473
The fitting parameters of the model: sample size=165、 $-2LL = 150.37$ 、 AIC=158.37					

Table 4. Regression model prediction accuracy

Actual Classification	**Forecast Classification**		**Total (rate)**
	Accident (rate)	**Non-accident (rate)**	
Non-accident	45 (34.09%)	87 (65.91%)	132 (100%)
Accident	20 (60.61%)	13 (39.39%)	33 (100%)

thereby increasing the risk of traffic accident; the vehicle speed at the location of the accident decreases quickly, and it propagates from downstream to upstream in the form of shock wave, therefore the speed of upstream coil will appear mutation, so *svavg* decreases, indicating that the risk of accident increases; when *qdavg* increases, that is the difference between the volume of large car and the volume of small car increases, easily causing the instability of traffic flow, so the risk of accident will increase too.

So, the change of *svavg*, *lqsd* and *qdavg* between 10 to 15 minutes prior to the time of the accident are the most effective to predict the probability of accident, which are used as the risk characterization factors of real-time safety assessment for the traffic flow operation.

The final model of the security situation deduction is as follows:

$$P(x_i) = \frac{1}{1+e^{-x_i'\beta}} = \frac{1}{1+e^{-(0.48457-0.03398svavg+0.40067lqsd-0.16874qdavg)}} \quad (6)$$

Where, *svavg*, *lqsd* and *qdavg* are the average value of speed of small car, the standard deviation of volume of large car and the average value of volume difference between large car and small car between 10 to 15 minutes prior to the time of the accident respectively.

After specifying a reasonable threshold value, the calibrated model can predict the risk of freeway traffic accidents in real-time. Because in the total samples, the proportion of accidents is 20%, so the threshold is set to 0.2 in this paper, namely when the probability of the model output is greater than 0.2, it is discriminated as the traffic accident; and when the probability of the model output is less than 0.2, it is discriminated as the safety state with non-accident. The model prediction accuracy is shown in Tab. 4.

As shown in Tab. 4, the Logistic model based on the traffic flow data of Beijing-Tianjin-Tanggu freeway can predict 60.61% of accidents and 65.91% of non-accidents, and the total prediction accuracy is 64.85%. Therefore, the accident risk prediction model established in this section can use the real-time traffic flow data to predict the risk of traffic accident of freeway well. In order to reduce the false-positive rate in actual applications, it can improve the prediction threshold value of the model according to the specific human and material resources allocation situation, such as increasing to 0.4 or 0.5.

5 Conclusions

The traffic flow data of the nearest detectors prior to or back to the location of the typical accidents in Beijing-Tianjin-Tanggu freeway are extracted in this paper, on the basis, we select the binary Logistic regression method to establish the security risk prediction model and realize the real-time prediction of the probability of accident by using the average value of speed of small car, the standard deviation of volume of large car and the average value of volume difference between large car and small car between 10 to 15 minutes prior to the time of the accident, which providing the more scientific support for the traffic control and the traffic emergency management decisions of freeway.

Acknowledgment

Fund: Applied Basic Research Project of Ministry of Transport (2013319223170).

References

1. Beijing Traffic Research Center. *TMC Beijing Live-presentation Meeting Information* **12**, 15-18 (Beijing, 2005).
2. G. Wang, X. Guo, Y. Jiang, J. Hou. Journal of Highway and Transportation Research and Development **27(4)**, 155-158 (2010).
3. Y. Liang, Y. Chen, F. Ren. Journal of Highway and Transportation Research and Development **22(12)**, 105-108 (2005).
4. K. Chen, W. Zou, X. Li. Highways & Automotive Applications **3**, 59-61 (2012).
5. J. Chen, W. Zhou. Journal of Chang'an University (Natural Science Edition) **22(4)**, 52-54 (2002).
6. H. Guo, H. Shi. Journal of Highway and Transportation Research and Development **22(8)**, 102-105 (2005).
7. Q. Xu, R. Yang. Journal of Highway and Transportation Research and Development **22(12)**, 131-134 (2005).
8. Weil R, Wootton J, Garcia-ortiz A. Mathematics Computer Modeling **27(9)**, 257 (1998).
9. Lee Chris Choongho. *School of Civil Engineering, The University of Waterloo*, (Ontario, 2004).
10. Abdelaty M, Uddin N, Abdalla F, et al. Transportation Research Record **1897**, 88-95 (2004).
11. Abdelaty M, Pande A. Journal of Safety Research **36(1)**, 97-108 (2005).
12. Abdelaty M, Uddin N, Pande A. Transportation Research Record **1908**, 51-58 (2005).
13. Lee C, Hellinga B, Saccomanno F. Transportation Research Record **2749**, 67-77 (2003).
14. Lee C, Saccomanno F, Hellinga B. Transportation Research Record **1784**, 1-8 (2002).
15. Hossain M, Muromachi Y. *The 89th Annual Meeting of the Transportation Research Board*, 1-15 (Washington, D. C., 2010)