Journal of Southeast University (English Edition) Vol. 31, No. 2, pp. 276 - 281 June 2015 ISSN 1003—7985

Construction of crash prediction model of freeway basic segment based on interactive influence of explanatory variables

Wang Xiaofei[1] Li Xinwei[1,2] Fu Xinsha[1] Zhao Lixuan[3] Liu Xiaofeng[4]

([1]School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510641, China)
([2]Guangzhou Expressway Co., Ltd, Guangzhou 510288, China)
([3]Guangdong Police College, Guangzhou 510230, China)
([4]Guangdong Traffic Group Co., Ltd, Guangzhou 510623, China)

Abstract: In order to improve the prediction precision of the safety performance function (SPF) of freeway basic segments, design and crash data of 640 segments are collected from different institutions. Three negative binomial (NB) regression models and three generalized negative binomial (GNB) regression models are built to prove that the interactive influence of explanatory variables plays an important role in fitting goodness. The effective use of the GNB model in analyzing the interactive influence of explanatory variables and predicting freeway basic segments is demonstrated. Among six models, the two models (one is the NB model and the other is the GNB model.) which consider the interactive influence of the annual average daily traffic (AADT) and length are more reasonable for predicting results. Furthermore, a comprehensive study is carried out to prove that when considering the interactive influence, the NB and GNB models have almost the same fitting performance in estimating the crashes, among which the GNB model is slightly better for prediction performance.

Key words: crash; freeway; safety performance function (SPF); interactive influence of explanatory variables; generalized negative binomial (GNB)

doi: 10.3969/j.issn.1003-7985.2015.02.021

Compared to other kinds of highways, the freeway is often designed with a relatively good driving environment, such as high alignment indices, a good pavement, being completely closed, no pedestrians, no interference from low speed, effective traffic safety devices, and so on. Thus, the accident rate and death toll on the freeway is within an average of 30% to 51% and 43% to 76% of ordinary highways in developed countries. However, the number of accidents, death tolls, injury tolls and the direct loss of property is 3.2, 8.4, 7.2 and 24.3 times the ordinary average for highways in China. Therefore, it is important to determine the actual circumstances of accidents occurring on freeways and how freeway environments influence the accident rates based on reliable databases.

Over the past several decades, historical surveys covering the characteristics and frequency of accidents involving freeways has been a very active research area[1-2]. However, in terms of freeway accidents in China, no specialized accident databases and highway design databases have yet been made available. Also, there is very little investigation clarifying China's current situation. Zhong et al.[3-4] developed the crash prediction model with a relatively small number of samples. Therefore, this paper attempts to establish models with large samples.

Mathematical statistics and regression analysis have been common methods used in predicting highway crashes. Other methods, such as fuzzy mathematics, the grey theory, the nerve cell method and clustering analysis, have also been used to establish the prediction models. However, freeway accidents are the results of the combined influence of multiple factors, such as alignment, traffic volume, the presence of an interchange or other structures. The above-mentioned methods explain how a single factor influences the crashes but fail to explain the interactions between these factors and how these influence the crashes. For this reason, when studying the crash prediction models, the freeway is often divided into several segments (a basic segment, general segment and special segment). The prediction function of the basic segment is also called the safety performance function (SPF) and it is the basis of the others.

The parameters of the SPF are the length of the segment and the traffic volume. The prediction result is the number of crashes. As for general segments or special segments, the crash number can be modified by crash modification factors (CMFs). Thus, the SPF is the basis of the freeway crash prediction model and the precision of the final result will be directly determined by the SPF. In order to determine the combined influence of multifactors, flexibility is introduced here to explain the influence. Flexibility is used in the manufacturing industry to explain the variational environment or the probabilistic

Received 2014-11-21.

Biography: Wang Xiaofei (1980—), female, doctor, lecturer, xiaofeiw@scut.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 51408229, 51278202), the Program of the Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University (No. K201204), the Science and Technology Program of Guangdong Communication Department (No. 2013-02-068).

Citation: Wang Xiaofei, Li Xinwei, Fu Xinsha, et al. Construction of crash prediction model of freeway basic segment based on interactive influence of explanatory variables [J]. Journal of Southeast University (English Edition), 2015, 31(2): 276 - 281. [doi: 10.3969/j.issn.1003-7985.2015.02.021]

ability from the variation[5]. The Cobb-Douglas production function, the linear production function, the Leontief production function, the variable elasticity of substitution (VES) production function and the transcendental logarithmic (Trans-log) production function are often used to analyze flexibility[6-8]. Among these methods, the Trans-log production function is the most popular function used to analyze traffic problems. Thus, the Trans-log function is adopted in this paper to study the difference between the situations with and without considering the combined influence of multifactors. The model with the best fitting degree is chosen as the SPF. Then the SPF is checked by the real traffic accident data.

1 Model Format and Basic Segment Definition

In this paper, the model format is as[9]

$$N_{e,x} = N_{SPF,x}\mathrm{CMF}_{AADT,x}\mathrm{CMF}_{lane,x}\cdots\mathrm{CMF}_{light,x} \quad (1)$$

where $N_{e,x}$ is the predictive model estimate of the crash number for a specific year on site type x; $N_{SPF,x}$ is the predicted average crash number determined by the SPF on site type x; $\mathrm{CMF}_{AADT,x}\cdots\mathrm{CMF}_{light,x}$ are the crash modification factors specific to site type x.

The basic segment for SPF is defined as follows.

Lane number: Two-way 4-lane;

Lane width: 3.75 m;

Hard shoulder: On both sides;

Median separator: Yes;

Crash barrier: On both sides;

Lighting: None;

AADT (two directions): No more than 5.76 (10^4 pcu/d);

Open to traffic duration: No less than two years and no reconstruction in two years.

2 Data

In order to acquire sufficient samples for a meaningful statistical analysis, six major sources are used: the National Statistics Annual Report of Road Traffic Accidents (NSARRTA,2013)[10], the Statistical Bulletin of Transportation Industry Development (SBTID, 2013), accident data from the Traffic Management Committee of Guangdong Province (MCGP), accident data from different Traffic Police Detachments (TPD, 7 freeways, 593.099 km in total), accident data from different Freeway Administrations and Maintenance Centers accordingly (FAMC, 7 freeways, 593.099 km in total), and additional results provided by other scholars. As for the sample size, see Tab.1.

Tab.1 Sample source and size

Source	Observation period	Freeway length/km	Accident amount
MCGP	2008 to 2012	2 200.779	135 498
NSG freeway TPD & FAMC	2008 to 2012	72.00	1 428
GZJC freeway TPD & FAMC	2008 to 2012	50.74	3 115
JZN freeway TPD & FAMC	2006 to 2012	109.84	11 209
GH freeway TPD & FAMC	2008 to 2012	155.306	3 441
KY freeway TPD & FAMC	2008 to 2012	125.20	2 351
GZBH freeway TPD & FAMC	2007 to 2012	21.652	12 850
SM freeway TPD & FAMC	2008 to 2011	58.361	719

3 Method

The basic function of the Trans-log NB production function is as follows[11-12]:

$$\ln Y = \alpha_0 + \alpha_k \ln K + \alpha_L \ln L + \alpha_{kk}(\ln K)^2 + \alpha_{ll}(\ln L)^2 + \alpha_{kl}\ln K \ln L$$

where Y is the dependent variable; K, L are the explanatory variables; and α_0, α_k, α_L, α_{kk}, α_{ll}, α_{kl} are the estimated parameters.

In this paper, the annual average daily traffic (AADT) Q_i and the segment length L_i are chosen as explanatory variables. If not taking the influence between these two variables into consideration, the NB function is as follows:

$$\ln \mu_i + \alpha_0 + \alpha_1 \ln(Q_i) + \alpha_2 \ln(L_i) \quad (2)$$

where μ_i is the estimate of crash number for a specific year of segment i; Q_i is the AADT for a specific year of segment i; $\alpha_k (k = 0, 1, \cdots, 5)$ are the estimated parameters.

If taking the influence between these two variables into account, the trans-log NB function is as follows:

$$\ln \mu_i = \alpha_0 + \alpha_1 \ln(Q_i) + \alpha_2 \ln(L_1) + \alpha_3[\ln(Q_i)]^2 + \alpha_4[\ln(L_i)]^2 + \alpha_5 \ln(Q_i)\ln(L_i) \quad (3)$$

Thus, six models are used as shown in Tab.2. Models 1 to 4 do not take the influence between two variables into account and Models 5 and 6 consider the influence. Models 1, 2 and 5 are NB regression models and others are GNB regression ones.

Tab.2 Models adopted and their estimated parameters

Model	Basic function form	Estimated parameters
Model 1 (NB)	$\ln\mu_i = \alpha_0 + \alpha_1 \ln(Q_i) + \alpha_2 \ln(L_i)$	$\alpha_0, \alpha_1, \alpha_2, \beta$
Model 2 (NB)	$\ln\mu_i = \alpha_0 + \alpha_1 \ln(Q_i) + \ln(L_i)$	$\alpha_0, \alpha_1, \beta$
Model 3 (GNB)	$\ln\mu_i = \alpha_0 + \alpha_1 \ln(Q_i) + \alpha_2 \ln(L_i)$, $\beta_i = e^{(\lambda_0 + \lambda_1 \ln L_i)}$	$\alpha_0, \alpha_1, \alpha_2, \lambda_0, \lambda_1 \lambda_{0\sim1}$
Model 4 (GNB)	$\ln\mu_i = \alpha_0 + \alpha_1 \ln(Q_i) + \ln(L_i)$, $\beta_i = e^{(\lambda_0 + \lambda_1 \ln L_i)}$	$\alpha_0, \alpha_1, \lambda_0, \lambda_1 \lambda_{0\sim1}$
Model 5 (NB)	$\ln\mu_i = \alpha_0 + \alpha_1 \ln(Q_i) + \alpha_2 \ln(L_i) + \alpha_3[\ln(Q_i)]^2 + \alpha_4[\ln(L_i)]^2 + \alpha_5 \ln(Q_i)\ln(L_i)$	$\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \beta$
Model 6 (GNB)	$\ln\mu_i = \alpha_0 + \alpha_1 \ln(Q_i) + \alpha_2 \ln(L_i) + \alpha_3[\ln(Q_i)]^2 + \alpha_4[\ln(L_i)]^2 + \alpha_5 \ln(Q_i)\ln(L_i)$, $\beta_i = e^{(\lambda_0 + \lambda_1 \ln L_i)}$	$\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \lambda_0, \lambda_1$

Note: β is the excessive dispersion coefficient of the NB or the GNB model.

Akaike information criterions (AIC)[13], Bayesian information criterions (BIC) and Pseudo R2 are adopted to test the accuracy of the prediction models. Cumulative residual and excessive dispersion coefficients are also used to evaluate the fit goodness of the models.

4 Analysis Process and Results

4.1 Models discussion

Estimated parameters which have been demarcated by survey data and parameters which reflect fit goodness have also been calculated. The results are listed in Tab. 3.

The results shown in Tab. 3 indicate that the interactive influence between two variables is clear. Thus, Model 5 and Model 6, which take the interactive influence into consideration, have better fitting, particularly compared to Model 2 and Model 4. However, the fitting goodness difference among Models 1, 3, 5 and 6 is not very obvious, so further analysis is carried out to determine which one is more accurate. In the scope of the definition condition, the crash number can be predicted and the estimated results are shown in Fig. 1 and Fig. 2.

Tab. 3 Estimated and statistical parameters

Estimated parameters	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
α_0	-0.11	-0.69	-0.10	-0.70	-0.13	-0.11
α_1	0.48	0.62	0.47	0.62	0.87	0.85
α_2	0.51	1.00	0.50	1.00	0.10	0.09
α_3					-0.23	-0.22
α_4					0.20	0.20
α_5					0.06	0.06
λ_0			-1.21	-0.76		-1.17
λ_1			0.16	-0.06		0.08
β	0.353	0.441			0.337 3	
AIC	2 482.864	2 540.805	2 484.392	2 542.730	2 479.306 0	2 481.200
BIC	2 483.182	2 541.043	2 484.790	2 543.048	2 479.863 0	2 481.835
Pseudo R^2	0.034	0.022	0.033	0.022	0.037 9	0.036
LR chi2	87.480	57.900	83.820	57.960	97.040	93.020

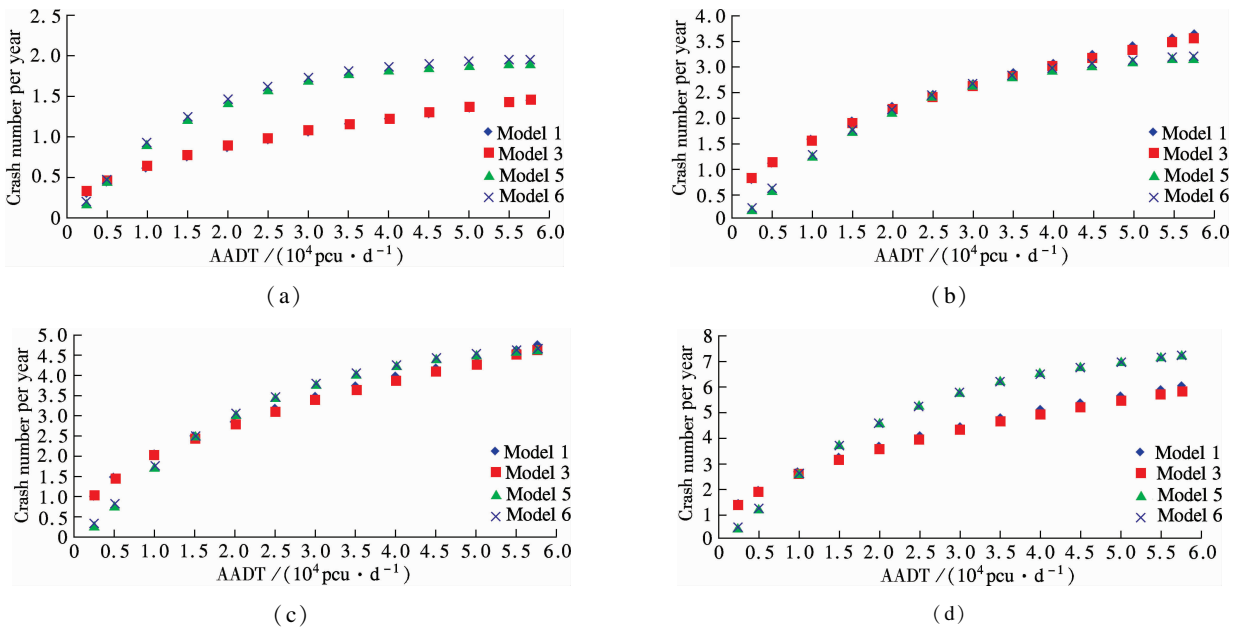
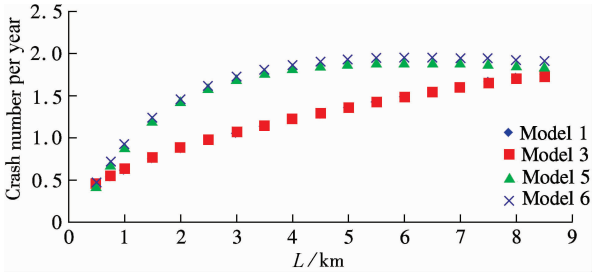


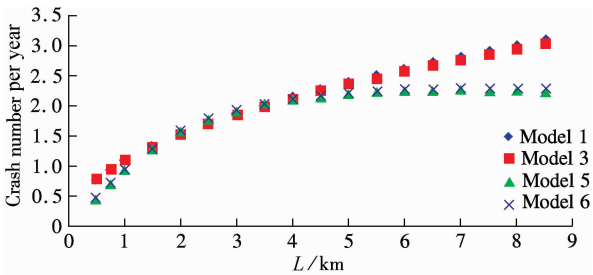
Fig. 1 Variation tendency of estimated results with AADT under different L. (a) L=0.5 km; (b) L=3.0 km; (c) L=5.0 km; (d) L=8.0 km

The results shown in Figs. 1 and 2 show no difference between Models 1 and 3 or Models 5 and 6. As for Models 1&3 and Models 5&6, the results reveal a remarkable difference.

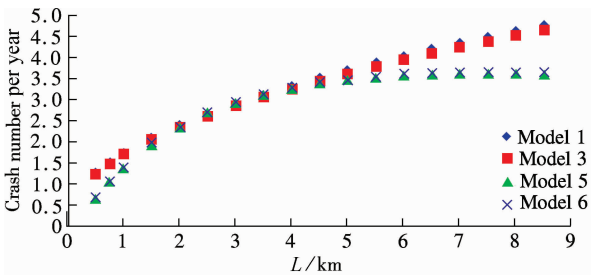
When the segment length is certain, crashes increase with AADT, as shown in Fig. 1. The increase tendency shows that Models 5 and 6 have a clear increase with AADT when the AADT is less than 3.5×10^4 pcu/d, however, the increase tendency is relatively gradual when AADT is more than 3.5×10^4 pcu/d. In comparison with Models 5 and 6, the increase tendency of Models 1 and 3 appears to have uniform variations. The survey data shows that there is one point after which the tendency changes. This may be due to the fact that with the increase in the traffic volume, speed and driving space decrease, resulting in fewer crashes. Thus, Models 5 and 6 reveal a more accurate picture.



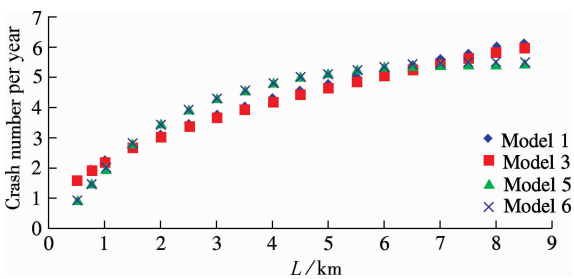
(a)



(b)



(c)



(d)

Fig. 2 Variation tendency of estimated results with segment length under different AADTs. (a) 0.5×10^4 pcu/d; (b) 1.5×10^4 pcu/d; (c) 3.5×10^4 pcu/d; (d) 5.76×10^4 pcu/d

The results shown in Fig. 2 with respect to segment length are not surprising. When AADT is certain, crashes increase with the segment length. The increase tendency also shows that Models 5 and 6 have a turning point at the length of 4.5 km. It reveals that the traffic flow tends to be steady in the same segment after several minutes' driving. So, the crashes will not increase as much as before.

In conclusion, Models 5 and 6 show a good corroboration with actual reality. Thus, the prediction models, Model 5 based on NB regression and Model 6 based on GNB regression have taken the interactive influence into consideration much better than others.

4.2 Discussion of NB Model 5 and GNB Model 6

To examine these two models in more detail and test their accuracy, the cumulative residual and excessive dispersion coefficients are introduced to evaluate these two models.

1) Cumulative residual

Cumulative residual can be calculated by

$$C_{\mathrm{R}}=\sum_{i=1}^{n}\frac{y_i-\hat{y}_i}{\sqrt{\hat{y}_i+\beta(\hat{y}_i)^2}} \tag{4}$$

where C_{R} is the cumulative residuals; y_i is the predicted values; $\hat{y}_i$ is the mean value of predicted results; n is the sample amount; β is the excessive dispersion coefficient.

Cumulative residuals will be centered at zero if the model fit is correct and the maximum threshold is the square root of the sample quantity, that is $(-25.298, 25.298)$. Cumulative residuals can be used to test the above NB and GNB models. The results in Fig. 3 show no difference between the two models. Most of the points are within the threshold scope and near the x-axis. Few residuals are less than the minimum threshold (-25.298). Thus, the two models are proved to be a good fit and show no difference between each other.

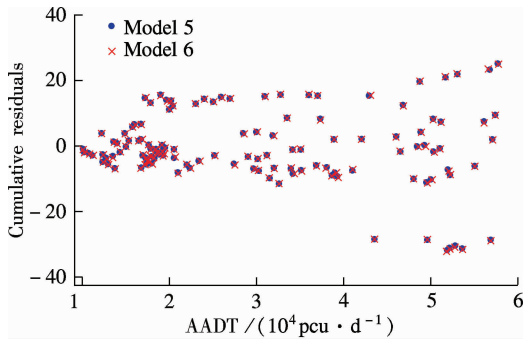


Fig. 3 Cumulative residuals of Models 5 and 6

2) Excessive dispersion coefficient β

The excessive dispersion coefficient of the NB model is constant and that of Model 5 is 0.337 3. As for the GNB, the excessive dispersion coefficient is the function of the explanatory variable $\ln L$, which can be expressed as

$$\beta=\mathrm{e}^{(\lambda_0+\lambda_1\ln L)} \tag{5}$$

where λ_0, λ_1 are the estimated parameters. In the example, $\lambda_0=-1.17$, $\lambda_1=0.08$. The regression results are shown in Tab. 4 and Fig. 4.

Further study discovers that the excessive dispersion

Tab. 4 Statistical index of the excessive dispersion coefficient of Model 6

Excessive dispersity	Sample quantity	Mean	Standard deviation	Minimum value	Maximum value
β	640	0.332 0	0.015 9	0.293 6	0.372 7

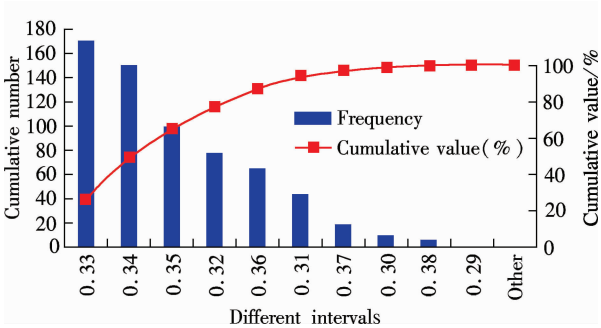


Fig. 4 Excessive dispersion coefficients at different interval distributions of Model 6

coefficient of Models 5 and 6 are almost equal. It shows that the mean dispersion coefficient of Model 6 (0.332 0) is slightly better than that of Model 5 (0.337 3). In conclusion, testing by cumulative residual and excessive dispersion, the two models demonstrate a good fit and there is almost no difference. They can both be used as the SPF of basic segment. In this paper, Model 6 (GNB) is adopted as SPF for further study.

4.3 Results

According to the above analysis, the following model is proposed as the freeway basic segment defined in Section 1:

$$N_{\mathrm{SPF},x}=\mathrm{e}^{[-0.11+0.85\ln \mathrm{AADT}_x+0.09\ln L_x-0.22(\ln \mathrm{AADT}_x)^2+0.20(\ln L_x)^2+0.06\ln \mathrm{AADT}_x\ln L_x]} \quad (6)$$

where AADT_x is the AADT of basic segment x, 10^4 pcu/d; L_x is the length of basic segment x, km.

In order to test the model, crashes estimated by Eq. (6) are compared with the actual data (see Tab. 5). From Tab. 5, it can be seen that the prediction crashes are very close to the actual values. It therefore proves that the SPF based on GNB can predict crashes correctly.

Tab. 5 Comparison of the crashes estimated by Eq. (6) with the actual data

L/km	AADT/ (10^4 pcu · d^{-1})	Predicted value	Actual value	Error
2.853	0.344 9	0.802	1	0.198
1.036	0.344 9	1.112	2	0.888
2.403	0.344 9	0.907	0	−0.907
0.601	0.344 9	0.789	1	0.211
0.734	0.344 9	0.766	1	0.234
1.520	0.344 9	0.759	1	0.241
0.616	0.344 9	0.875	1	0.125
0.697	0.344 9	0.771	0	−0.771
0.775	0.344 9	0.822	1	0.178
0.736	0.344 9	0.759	1	0.241
1.408	0.180 0	0.415	0	−0.415
Std error		0.540 0	2.000	0.156
Max		1.112 0	2.000	−0.907
Total		16.837 0	15.000	

5 Conclusion

The analysis sheds light on the safety performance function (SPF) of the freeway basic segment. With detailed analysis and study, some conclusions are drawn as follows:

1) With enough samples and data, the effective use of the GNB model in analyzing the interactive influence of explanatory variables and predicting crashes on the freeway basic segment has been proved.

2) The contribution of interactive influence between the NB model and the GNB model is compared. The results show that when interactive influence is taken into consideration, the prediction results of the crash increase tendency becomes more accurate by using AADT or the length.

3) Furthermore, comprehensive study proves that when considering the interactive influence, the NB and GNB models have almost the same good fit when estimating the crashes, among which the GNB model is slightly better.

References

[1] Durduran S S. A decision making system to automatic recognize of traffic accidents on the basis of a GIS platform[J]. *Expert Systems with Applications*, 2010, **37**(12): 7729−7736.

[2] White J, Thompson C, Turner H. WreckWatch: automatic traffic accident and notification with smartphones [J]. *Mobile Networks and Applications*, 2011, **16**(3): 285−303.

[3] Zhong Liande. Research on accident prediction model of freeway [D]. Beijing, China: Key Laboratory of Traffic Engineering, Beijing University of Technology, 2008. (in Chinese)

[4] Ma Zhuanglin. Temporal-spatial analysis model of traffic accident and its prevention method on expressway [D]. Beijing, China: School of Traffic and Transportation, Beijing Jiaotong University, 2010. (in Chinese)

[5] Mandalbaum M. Flexibility in decision making: an exploration and unification [D]. Toronto, Canada: Department of Engineering, University of Toronto, 1978.

[6] Wu Lurong, Liang Fangfang, Shi Zhikai. China's translog production function model of losses of traffic accidents[J]. *Mathematics in Practice and Theory*, 2010, **40**(22): 56−61.

[7] Fridstrom L, Ifver J, Ingebrigtsen S, et al. Measuring the contribution of randomness, exposure, weather and daylight to the variation in road accident counts[J]. *Accident Analysis and Prevention*, 1995, **27**(1): 1−20.

[8] Zeng Juan. Research on influencing factors of socio-economic losses of road traffic accident based on generalized linear models[J]. *Journal of Wuhan University of Technology*, 2010, **32**(6): 155−158. (in Chinese)

[10] Ministry of Transport of the People's Republic of China. Statistical bulletin of communication and transportation industry development in 2013 [EB/OL]. (2014-05-13) [2014-06-20]. http://www.moc.gov.cn/zfxxgk/bns-

sj/zhghs/201405/t20140513_1618277.html. (in Chinese)
[11] Christensen L R, Jorgenson D W, Lau L J. Transcendental logarithmic production frontiers[J]. *Review of Economics & Statistics*, 1973, **55**(1):28-45.
[12] Li Rong, Liu Xiang, Liu Jian. Research on traffic accident frequency prediction based on translog production function[J]. *Journal of Hunan University: Natural Sciences*, 2013, **40**(4):49-54. (in Chinese)
[13] Bozdogan H. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions[J]. *Psychometrika*, 1987, **52**(3):345-370.

基于变量相互影响的高速公路基本路段事故预测模型构建方法

王晓飞[1] 李新伟[1,2] 符锌砂[1] 赵立莹[3] 刘小峰[4]

([1] 华南理工大学土木与交通学院,广州 510641)
([2] 广州市高速公路有限公司,广州 510288)
([3] 广东省警官学院,广州 510230)
([4] 广东省交通集团有限公司,广州 510623)

摘要: 为了提高基本路段事故预测模型(SPF)的预测精度,收集了640个基本路段设计资料及事故资料,应用3个负二项回归模型(NB)和3个广义负二项(GNB)回归模型对收集的数据进行拟合,并分析了解释变量的交互影响.研究表明在上述6个模型中,其中考虑了年平均日交通量和路段长度交互影响的2个模型(一个为NB,另一个为GNB),其预测结果更为合理.进一步综合对比表明考虑交互影响时,NB模型和GNB模型的适用性几乎相同,而GNB略佳.

关键词: 事故;高速公路;事故预测模型;解释变量交互影响;广义负二项模型
中图分类号: U412.3