Modelling highway safety with data from Chinese freeways

X. Sun[1,2], Y. He[1], L. Zhong[1] & Y. Chen[1]
[1]*Beijing University of Technology, China*
[2]*University of Louisiana, USA*

Abstract

This paper presents a study on the development of safety performance function for freeways. Applying well-established statistical methods, we evaluated all variables that may affect freeway safety and selected the most significant ones in the model. A variable analysis unit was utilized in this study to overcome the difficulties in obtaining accurate crash and highway attribute data, as well as to improve the modelling quality. The results of this study provide much needed tools for freeway safety analysis.

Keywords: freeway safety, safety performance model, probability distribution.

1 Introduction

The economic boom in China over the past two decades has stimulated unprecedented freeway construction. In the last 20 years, the freeway mileage in the country has increased from zero to 53,900 kilometres. The freeway network has greatly enhanced the capacity of the national highway transportation system, further fuelling economic development, and significantly changing the lifestyle of ordinary Chinese people. The planned 85,000 kilometre freeway network is centred in Beijing, the capital of China, with 7 radial lines, 9 north-south lines and 18 east-west lines, which enables people to access a freeway in about 30 minutes in the Eastern Region, 60 minutes in the Central Area, and 2 hours in the Western Territory.

However, this "overnight" success also comes at a human cost. Traffic crashes on the freeways have increased rapidly. The safety of freeways has become a big concern to both the travelling public and the highway agencies in the country. This safest type of highway has been perceived as the most

WIT Transactions on the Built Environment, Vol 108, © 2009 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/SAFE090441

dangerous highway by many motorists. About 6,600 people died on freeways in 2006, which represents 8.3% of all highway traffic fatalities based on the published official statistics [1]. After 20 years of freeway construction and operation, sustainability is becoming an overarching design principle for freeway systems with safety as one of the important aspects in sustainable transportation development.

To reduce freeway crashes, it is crucial to have a qualitative safety evaluation tool. A safety performance model provides such a tool that can be used to assess a particular freeway's safety level and help identify effective crash countermeasures. A well established and validated safety performance model can be used in observational before-and-after studies, and in network screening for safety management systems. The unique characteristics of freeway crashes in China, as discussed in our previous paper, indicate the need to model freeway safety independently.

It is easier to develop a SPF for the lower class of highways, because they bear lower geometric design features. Highway segments with lower geometric design features, such as small horizontal curve radiuses, narrow lane and shoulder width, and steeper grade of vertical curves, have been long recognized as vulnerable locations for traffic crashes, thus easier to be modelled. Both the crash rate (crashes per million vehicle mile/km travelled), and the fatal crash rate (fatalities per 100 million vehicle mile/km driven) for freeways are generally much lower than that of other types of highways. The difficulties in linking safety with design features are the main reason why fewer studies have been conducted.

2 Literature review

Because of its importance in traffic safety evaluation, many studies have been conducted in modelling highway safety performance [2–5]. Selecting variables that significantly affect safety is one of the key steps in developing a valid SPF. Several papers discuss how to select modelling variables in detail [6–8]. Highway general features, such as horizontal and vertical arguments and the length of down/uphill segments, are recognized as important variables by these studies.

A study specifically investigated the analysis unit [9], another important aspect in SPF development. This study proposed a new method of freeway section division based on the ordinal clustering method, which uses the clustering index expressed as crash frequency per kilometre. The advantages of this method include: easy identification of high crash locations, easily associating safety with a host of selected variables, and better exploration of crash probability distributions.

The generalized linear regression method was widely used by many previous studies [9–14] over the past 10 years, which reflected the progress on the crash modelling. The method assumed that crashes occurring on a particular roadway are independent, stochastic events, which follow certain probability distributions. Poisson and Negative Binomial (NB) models are well-accepted methods of modelling the crashes. Zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) regression models have recently been applied in safety modelling.

3 Methodology

3.1 Data

There are two types of data required for modelling highway safety: crash data and highway attributes data by analysis unit. Crash data and highway attributes data from more than 10 freeways were collected for this study. Similar to other developing countries, obtaining reliable crash data is a big challenge considering somewhat inconsistent data recording practices around the country.

Not all data were used in this model development and validation. The freeway with the most complete crash data and highway attribute data was used in model development. This 142 kilometre four-lane freeway (2 lanes in each direction) was one of the very first freeways built in China. The details of 2829 reported crashes were collected and compiled. Each crash record contains information on the occupants' demographics, vehicle characteristics, environmental conditions, and crash information, such as crash severity, time, location and type.

To overcome the small data size problem, this study used hour instead of year as the analysis time unit. That is, hourly flow and hourly crash data were used in the modelling process. The initial analysis revealed a close relationship between the distribution of hourly crashes and hourly traffic volume. Traffic information, average speed, vehicle type and occupancy were obtained from 23 cross-sections spaced approximately 6 kilometres apart in both directions at every minute.

Table 1: Summary statistics of segment length in kilometres.

Average	Std. Deviation	Min.	Max.
1.86	1.34	1	8

3.2 Spatial analysis unit

Generally there are two methods in the selection of a spatial analysis unit: fixed length and variable length. The fixed length method divides a roadway into sections with uniform length, and the variable length method divides a roadway into sections with different length based on the change of highway attributes. These attributes are traffic volume, speed limit and geometric design elements, such as horizontal curvature, vertical grade, number of lanes, shoulder type and width, median type of width, type of pavement, and etc. Considering relatively small variations in freeway design elements and the problems with both methods, this study applied a new method based on the ordinal clustering process. By using the clustering index, expressed as crash frequency per

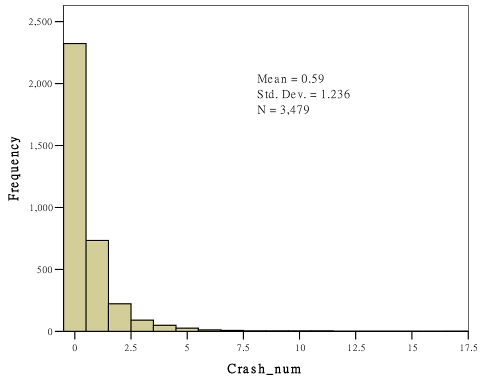

Figure 1: Distribution of hourly crashes.

kilometre, it is easier to identify high crash locations, associate safety with a host of selected variables, and explore crash probability distributions as described by [4].

Table 1 lists the characteristics of the segments defined by the clustering process. Figure 1 shows the distribution of crashes by hour from these segments.

3.3 Model development

3.3.1 Model format

A generalized linear regression method was developed in this study. As shown in Figure 1, the distribution of crash frequency is very much in line with Poisson and Negative Binomial distributions. According to the Kolmogorov-Smirnov test, the hourly crash distribution closely follows the negative binomial distribution with 95% confidence. Based on the preliminary statistical analysis, segment length and traffic volume are the most influential factors. Thus, the first model is established as:

$$\lambda_{ij} = EXPO \cdot Exp(\beta_0 + \sum_{k=1}^{n} \beta_{ij} x_{ijk}) \tag{1}$$

where:

λ_{ij} = predicted annual crash frequency to roadway segment i at the jth hour (24 hours in total);

$EXPO$ = an exposure variable expressed as millions of vehicle-kilometres;

β_{ij} = parameter;

x_{ijk} = explanatory variables for roadway segment i at the jth hour.

Hourly traffic volume correlates well with hourly crash counts. Most importantly, there is a clear variation in truck volume by hour of the day, which is associated with the time limit for trucks in the city. As mentioned previously, the time analysis unit is hour. This also helps to capture the impact of traffic and traffic composition on crash occurrences. Poisson, Negative Binomial, Zero Inflated Probability (ZIP), and Zero Inflated Probability Negative Binomial (ZINB) models were used for the distribution of hourly crashes separately.

3.3.2 Variable selection

Selecting statistically significant variables which are obtainable and quantifiable in practice is very important in developing a reliable and theoretically sound safety predictive model. Based on the unique characteristics of freeway traffic crashes in China and preliminary statistical analysis, variables of location, direction of travel, horizontal curvature, interchange, vertical grade, as well as traffic composition were selected.

Location was defined as either rural or city, since there is a significant difference in crash frequencies between these two areas. The time restriction on when truck traffic can begin entering the city prompted a variable of direction of travel. Freeway interchanges are changeling locations due to weaving traffic. It is particularly true in China, because of somewhat low design standards used in interchange geometrics, such as sharp curves for exit ramp, steep vertical slope, and short acceleration and deceleration lanes.

Table 2: Independent variables.

	Variable explanation	Variable Name	Value
Exposure	Exposure	EXPO	10^6 annual veh-km
Environment	Direction	Direction	0: outbound 1: inbound
	Location	City or rural	0: rural, 1: city
	Interchange area	Interchange	0: no, 1: yes
	Average angle of horizontal curves	Ave_angle	numerical
	Vertical variables	Ave_slope	numerical
Traffic variables	Percent of truck	Truck%	numerical
	Speed difference between car and truck	Spe_difference	numerical
	Standard deviation of truck speed	Spe_stan_truck	numerical
	Standard deviation of car speed	Spe_stan_car	numerical

Horizontal and vertical alignment variables were defined to capture the impact of horizontal curves including average angle of horizontal curve and average slope of a segment. Traffic variables are percentage of large vehicles, average speed of large vehicles and cars, and standard deviation of average speed for the two types of vehicles since our previous study showed these variables are significantly affecting highway safety, particularly on the frequency of rear-end collisions. Table 2 lists all variables considered first for the modelling.

3.3.3 Model development

To best capture the probabilistic nature of crashes, four different probability distribution modes were evaluated. They are Negative Binomial (NB), Poisson, ZIP, and ZINB. The selection of variables in each model went backward, i.e., firstly all the independent variables in Table 2 were used, and then the least irrelevant variable was rejected one at a time according to the output criterion (e.g. the P-value). Table 3 gives the regression outputs where K is the overdispersion parameter.

Table 4 lists the goodness of fit of the four models: Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), logarithm likelihood ratio test, and Vuong statistics. The smaller the value of AIC (or BIC), the better the model is. The best model can be selected by comparing the output measures. As shown in Table 4, all the numbers indicates that NB capture the crash data the best. The Chibar2 of 417.73 rejects the null hypothesis of no overdispersion. The Vuong test, 4.13>1.96, suggests that the ZIP is preferred to the Poisson.

Table 3: Summary of negative binomial regression.

Variable	Coefficient	Std. Error	z	p>!Z!	95% confidence
City_rural	1.119211	0.0762063	14.69	0.000	0.9698493
Interchange	0.4733442	0.0818434	5.78	0.000	0.3129342
Ave_angle	0.0112781	0.0029372	3.84	0.000	0.0055213
Truck%	1.375432	0.1322295	10.40	0.000	1.116267
Spe_stan_truck	0.0588855	0.0100452	5.86	0.000	0.0391972
Constant	-2.737629	0.160209	-17.09	0.000	-3.051633
K	0.9109513	0.0761533			0.7732801

Table 4: Model performance comparison.

Models	Log likelihood	AIC	BIC	Vuong	Chibar2
NB	-3115.2895	6225.348	390.506		417.73
Poisson	-3288.8923	6593.785	882.535		
ZIP	-3224.438	6466.877	669.576	4.30	
ZINB	-3104.683	6227.367	558.506	-0.07	

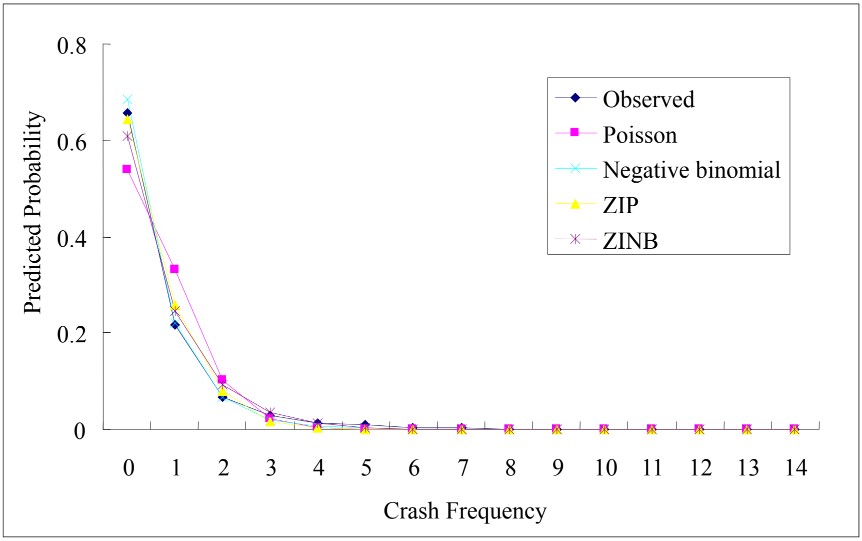

Figure 2: Model performance comparison.

Figure 2 gives the comparison of the predicted probability from all four models with the observed probability. It is clear that the NB model fits the data better than other models, and the Poisson model is the worst.

Finally the Negative binomial distribution was selected to present the probability of crash occurrence, and the best SPF was developed in the following function form:

$$\lambda_{ij} = EXPO \cdot EXP(-2.737629 + 1.119211 \cdot City\_rural + 0.4733442 \cdot Interchange$$

$$+0.0112781 * Ave.angle + 1.3754 * Truck\% + 0.0588885 * Spe\_s\tan\_tuck)$$

where

K=0.9109

λ_{ij} = predicted annual total crash frequency in roadway segment i at the jth hour.
Other variables explanations are given in Table 2.

4 Model validation

To further test the model robustness, another analysis called Cumulative Scaled Residuals (CURE) was applied. CURE is a useful tool for checking and adjusting the model fit. In general, a good CURE plot is one that oscillates around 0. A bad CURE plot is one that is entirely above or below 0 (except at the edges).

The CURE value is calculated as follows:

$$CURE = \sum_{i,j:x_{ij} \le l} \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\hat{y}_{ij} + K\left(\hat{y}_{ij}\right)^2}} \qquad (2)$$

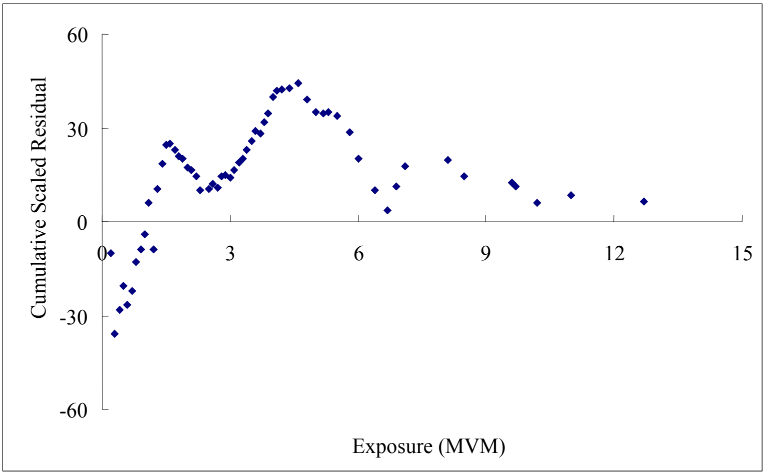

Figure 3: Cumulative scaled residual versus EXPO.

where

K: the overdispersion parameter

y_{ij} : observed crash count for highway segment no. i at the jth hour

$\hat{y}_{ij}$: estimated crash count for highway segment no. i at the jth hour

l: the range of X_{ij}

The Cumulative Scaled Residuals were plotted against leading explanatory variables for SPF, in which the CURE varies from -36 to 44.2, which is in the acceptable range (the threshold value is within ±56.7, which is approximately the square root of the number of data sets that is 3,214). It is somewhat clear that the model overestimates crash frequency when EXPO is larger. The Cumulative Scaled Residual analysis for other variables shows similar results.

A different data set that was not utilized in the model development was used for model validation. This four-lane freeway is 224km long with a standard cross section. There were 448 crashes occurrence in 14 months. The freeway was divided into 200 segments for both directions according to the Ordinal Clustering method. The summation of the predicted crash counts of each roadway segment is 489, just 41 or 9.1% higher than the observed 448.

5 Discussions

This paper presents the first attempt in modelling freeway safety performance in China. The results from the model development and validation indicate a very satisfactory precision. The precision can also be improved with Empirical Bayes (EB) procedure when crash data is available.

Due to the unique crash characteristics in China, four variables; location (City vs. Rural), presence of interchange, average speed and speed difference were selected in the final model. In the future as Chinese become more familiar with freeway operation and vehicles' operating capacities become more uniform, the situation could change. New variables could merge as important freeway safety influential factors, and the ones introduced in this paper, could become insignificant in a safety performance prediction model. At present time, the model developed in this paper would serve as a tool in freeway safety evaluation.

Acknowledgement

This study was partially sponsored by a research project under the Ministry of Communication of China.

References

[1] "The Annual Crash Report by Ministry of Security of China" 2006

[2] Golob, T. F., Recker, W W. A method for relating type of crash to traffic flow characteristics on urban freeways. Transportation Research Part A, 2004, 38(1): PP 53-80.

[3] Persaud, B. N. Estimating accident potential of Ontario road sections, Transportation Research Record 1327, Transportation Research Board, National Research Council, Washington, D.C., 1991: PP47~53

[4] Garber, N. J. and Ehrhart, A. The effect of speed, flow, and geometric characteristics on crash rates for different types of Virginia highways. Final Report, Virginia Transportation Research Council, Charlottesville, Virginia, January 1991

[5] Lord, D., Manar, A., Vizioli, A. Modelling Crash-Flow-Density and Crash-Flow-V/C Ration Relationships for Rural and Urban Freeway Segments. Presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2004

[6] Vogt, A., Bared, G. Accident models for two-lane roads: segments and intersections. Publication FHWA-RD-98-133. Federal Highway Administration, US. 1998. PP

[7] Chen, Y. S., Sun X. D., Zhou L. D., Zhang G. W., Speed Difference and Its Impact on Traffic Safety of One Freeway in China. Journal of Transportation Research Record, 2007

[8] Zhong, L. D., Sun, X. D., Chen Y. S., Zhang, G J., Exploring the Relationships between Crash Rates and Average Speed Differentia between Cars and Large Vehicles on a Suburban Freeway. Proceedings of IEEE ITSC 2006.Toronto, Canada. 2006:PP 1638~1641

[9] Jovanis, P. P. and Chang, H. L. Modelling the relationship of accidents to miles travelled, Transportation Research Record 1068, Transportation Research Board, National Research Council, Washington, D.C., 1986: 42~51

[10] Miaou, S. and Lum, H. Modelling vehicle accident and highway geometric design relationships, Accidents Analysis and Prevention, Vol. 25, No. 6, 1993: 689~709

[11] Hauer, E., Ng, J.C.N., Lovell, J. Estimation of safety at signalized intersections, Transportation Research Record 1185, Transportation Research Board, National Research Council, Washington, D.C., 1988: 48~61

[12] Hinde, J. and Demetrio, C. G. B. Overdispersion: models and estimation, Computational Statistics & Data Analysis Vol. 27, No. 2, 1998: 151~170

[13] Kumala, R. and Roine, M. Accident prediction models for two-lane roads in Finland, Proceedings of the Conference on Traffic Safety Theory and Research Methods. Amsterdam: SWOV, 1988: 48~61

[14] Kumala, R. Safety at rural three and four-arm junctions: development of accident prediction models, Espoo. Technical Research Centre of Finland, VTT Publications PP 233