Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Application of generalized estimating equations for crash frequency modeling with temporal correlation*

Wen-qing WU[1,2,3], Wei WANG[†‡1,2,3], Zhi-bin LI[1,2,3], Pan LIU[1,2,3], Yong WANG[4]

([1]Jiangsu Key Laboratory of Urban Intelligent Transportation Systems, Southeast University, Nanjing 210096, China)
([2]Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Nanjing 210096, China)
([3]School of Transportation, Southeast University, Nanjing 210096, China)
([4]School of Management, Chongqing Jiaotong University, Chongqing 400074, China)
[†]E-mail: wangwei@seu.edu.cn

Received Oct. 22, 2013; Revision accepted Apr. 15, 2014; Crosschecked June 24, 2014

Abstract: Traditional crash frequency modeling uses crash frequency data averaged across multiple years. When data size is small, crash data in each year are used in the modeling to extend the size of the samples. The extension of sample size could create a temporal correlation among crash frequencies of the different years, which could affect the modeling accuracy. The primary objective of this study is to evaluate the application of the generalized estimating equation (GEE) procedures to account for the temporal correlation in the longitudinal crash frequency data. Four-year crash data at exit ramps on a freeway in China were collected for modeling. Based on the same data, traditional generalized linear models (GLMs) were estimated for model comparison. Results showed that traditional GLM underestimated the standard errors of coefficients for explanatory variables. The GEE procedure with an exchangeable correlation structure successively captured the temporal correlation among the crash frequencies of the different years. The GLM with GEE outperformed the traditional GLM in providing a good fit for the crash frequency data. Results of this study can help researchers better understand how various factors affect the crash frequencies at freeway divergent areas and propose effective countermeasures.

Key words: Crash frequency, Generalized estimating equation (GEE), Temporal correlation, Freeway, Safety
doi:10.1631/jzus.A1300342 **Document code:** A **CLC number:** U491.31

1 Introduction

Road infrastructure and traffic volume have dramatically increased over the past decades in many countries, resulting in a considerable increase in traffic crashes (Saenghaengtham and Kanongchaiyos, 2006; Hu *et al.*, 2008; Jin *et al.*, 2011; Luoma and Sivak, 2012; Li *et al.*, 2012; 2013; 2014; Ma *et al.*, 2012; Jin *et al.*, 2013). An analysis on crash frequency can identify factors that affect the crashes to help reduce the number of crashes. Previously, numerous crash prediction models have been developed. Various methodologies have been proposed for crash frequency modeling to improve the predictive accuracy for crashes (Lord and Mannering, 2010).

In most of the previous crash prediction models, crash counts were usually aggregated over several years and the average crash count per year was considered the response variable (Bauer and Harwood, 1998; Bared *et al.*, 1999; McCartt *et al.*, 2004; Chen *et al.*, 2009; 2011; Lord and Mannering, 2010; Liu *et al.*, 2010; Washington *et al.*, 2010). The reason for this is to reduce the random variation in the yearly

‡ Corresponding author

* Project supported by the National Key Basic Research Program (973) of China (No. 2012CB725400), the National High-Tech R&D Program (863) of China (No. 2012AA112304), and the National Natural Science Foundation of China (No. 51338003)

crash data. The maximum likelihood estimation (MLE) method is used to estimate the coefficient and significance level of each predicting factor in the model. The MLE method can produce accurate model estimates when the dataset contains a large number of recordings.

However, researchers or agencies in many countries, especially in developing countries, are often faced with the small sample size issue in the crash data. Due to the restrictions on the crash reporting systems, there is usually no gateway for the public to access the resources of crash recordings as well as road and traffic information. Moreover, important information is often missing in the original dataset which further reduces the sample size that is useful for modeling. With the data of a small sample size, the desirable properties of some parameter-estimation techniques, such as the MLE, are not realized (Washington *et al.*, 2010). Biased estimates and incorrect inferences could occur in the crash prediction models.

To enlarge the sample size in the dataset, a natural consideration is to divide the data aggregated over several years into smaller time intervals (a unit of year) and treat the crash counts in each year as separate observations. The enlarged sample size would improve the estimating accuracy of the crash prediction model. However, disaggregating the crash data could create a temporal correlation in the dataset. Crash counts in different years could be correlated with each other due to the unobserved or unconsidered effects of factors associated with a specific road entity that do not change over years. This fact becomes rather determinate in developing countries since the information of important factors is often missing. The temporal correlation could adversely affect the precision of parameter estimates in the crash prediction model if not properly considered in the modeling procedure (Lord and Persaud, 2000).

Previously, safety researchers have proposed numerous sophisticated methodologies to account for the correlations among groups of crash frequencies. Those methodologies include the generalized estimating equation (GEE) (Lord and Persaud, 2000; Wang and Abdel-Aty, 2006; 2008), the random-effects model (Shankar *et al.*, 1998; Miaou and Lord, 2003; Quddus, 2008), the hierarchica/multilevel model (Jones and Jørgensen, 2003; Kim *et al.*, 2007), and the multivariate modeling approaches (Ma and Kockelman, 2006; Park and Lord, 2007; El-Basyouny and Sayed, 2009). However, several models may not be appropriate for practical applications because they are very complicated and difficult to solve. The safety researchers or agencies often experience great difficulties in trying to select the appropriate method for their particular needs.

In this study, the main objective is to evaluate the application of GEEs to account for the temporal correlation in the crash frequency modeling. To achieve the research objective, four-year crash data were collected from exit ramps on a freeway in China. Procedures of traditional generalized linear models (GLMs) as well as a GLM with GEE were estimated for model comparison. The findings of this study can provide useful information to researchers in developing crash prediction models based on the crash frequency data with temporal correlations.

2 Data

The data were collected from 32 sections of exit ramps on the Guangshen freeway in China. The freeway has a total length of 98 km and is located in the southern part of China. The freeway connects several of the most economically developed cities in Guangdong province and there is a high traffic demand on both the mainline and ramps. Traffic crashes frequently occur at the exit ramp areas on this freeway.

In this study, the exit ramp areas include two sub-segments that are located in the upstream and downstream of the painted nose. In previous studies, a section with 457.5 m (1500 ft) in the upstream and 304.8 m (1000 ft) in the downstream was considered as the influencing area of an exit ramp (Chen *et al.*, 2009; Liu *et al.*, 2010). On the Guangshen freeway, the message signs for exits are usually posted 500 m upstream of the exits. Thus, a section with 500 m in the upstream and 300 m in the downstream of the painted nose is considered as the exit ramp area. The illustration of exit ramp areas is shown in Fig. 1.

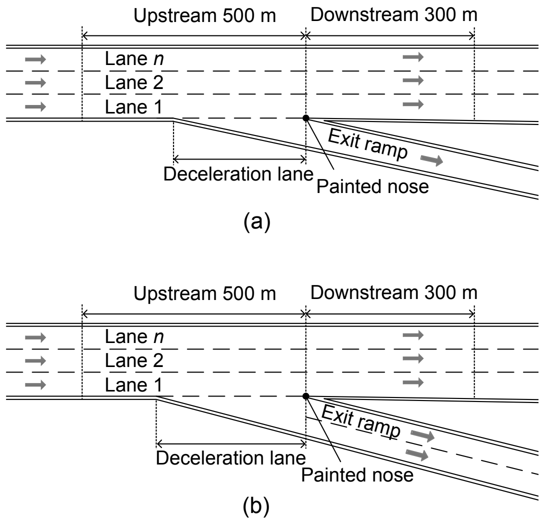

Fig. 1 Illustration of two types of exit ramps
(a) One-lane exit ramp; (b) Two-lane exit ramp

Local freeway agency would like to identify which factors are related to the number of crashes to evaluate the safety performance of exit ramps. They want to identify the ramps that have higher than normal crashes and implement countermeasures to reduce the crashes. To reach their objective, a large dataset is usually required to obtain accurate estimates on the safety impacts of explanatory factors (Hauer, 1997). However, it is difficult to obtain crash data as well as road and traffic information on other neighboring freeways since this information is not shared. Only the data from 32 exit ramps on the freeway are available.

Four-year crash data, from 2006 to 2009, were obtained from the local freeway management agency. A total of 4429 crashes were observed. The crashes included all types of injury severities. The statistics of the crash counts are shown in Table 1. The average crash count per year is 34.60 with a standard deviation (SD) of 38.87. The crash data have an obvious feature of over-dispersion since their mean values are much smaller than the variance.

There are two types of exit ramps that are typical on the Guangshen freeway, according to the number of exit lanes shown in Fig. 1. The road geometric attributes, such as the number of lanes, presence of longitudinal grade, and right shoulder width, are identified from the design drawing manual of the freeway. The speed limit information and average daily traffic of the mainline and exit ramps were obtained from the freeway management company. The percentage of days with severe weather in each year was also obtained from recordings. The summary statistics of these explanatory variables are shown in Table 2.

3 Hardware and embedded safe operation system (ES-OS)

This section briefly describes the traditional GLM for crash frequency modeling and the GEE procedure to account for the temporal correlation in longitudinal data. The cumulative residual test and type III analysis for model assessment are also introduced.

3.1 GLM for crash frequency analysis

When applying the GLM for crash frequency analysis, the random component is often likely to follow a Poisson or negative binomial (NB) distribution (Washington *et al.*, 2010). After reviewing the model specifications for crash frequency analysis at freeway exit ramps in (Lord and Mannering, 2010), the following model form is considered:

$$\begin{aligned}\ln(E\{u_t\})=&\ln\beta_0+\beta_1\ln(F_1(t))+\beta_2\ln(F_2(t))\\&+\beta_3X_3(t)+\ldots+\beta_JX_J(t),\end{aligned}\qquad(1)$$

where $\ln(E\{\mu_t\})$ is the natural log of expected crash frequency in period t at the exit ramps, u_t is the crash

Table 1 Summary statistics of dependent variables (sample size is 32)

Dependent variable	Total	Min	Max	Mean	SD
Crash count in 2006	1124	5	151	35.13	33.89
Crash count in 2007	1253	6	219	39.16	46.06
Crash count in 2008	1182	6	186	36.94	46.02
Crash count in 2009	870	1	120	27.19	26.80
Average crash count per year	1107	1	219	34.60	38.87

Table 2 Summary statistics of explanatory variables

Variable	Min	Max	Mean	SD	Frequency
Number of lanes on mainline	3	4	3.06	0.35	128
Number of lanes on exit ramp	1	2	1.94	0.24	128
Length of deceleration lane (km)	0.07	0.51	0.29	0.11	128
AADT* on exit ramp (in hundreds)	28	1127	243	195	128
AADT on mainline (in hundreds)	702	2831	1521	404	128
Difference of speed limits between mainline and exit ramps (km/h)	40	80	60.63	7.91	128
Right shoulder width (m)	1.03	3.84	2.90	0.76	128
Bad weather ratio	0.12	0.17	0.14	0.36	128
Lane balanced					128
0 (balanced)					124 (96.88%)**
1 (unbalanced)					4 (3.12%)
Grade					128
0 (no grade)					104 (81.25%)
1 (up/down grade)					24 (18.75%)
Land use					128
0 (rural)					100 (78.12%)
1 (urban)					28 (21.88%)

* AADT: annual average daily traffic; ** the values in brackets are percentages of samples

frequency in period t, $F_1(t)$ and $F_2(t)$ are annual average daily traffic (AADT) on the mainline and ramps in period t, respectively, $X_j(t)$ is the jth explanatory variable in period t, and β_j is the jth coefficient to be estimated (j=0, 1, …, J), J is the number of coefficient of variables. The average crash frequency per year across the four years was used in the GLM when t was set to 1 year. To enlarge the sample size, the crash frequency data could be disaggregated by a small time interval, which is one year at each ramp. Thus, the model based on yearly aggregated data is determined when t was set to 4 year.

3.2 GLM with GEE procedure

In the model based on yearly disaggregated data in Eq. (1), the crash frequency in a year could be correlated with others. The GEE is an extension of the GLM for estimating the temporally correlated data. Using the link function shown in Eq. (1), the coefficients $\boldsymbol{\beta}$ are estimated by (Lord and Persaud, 2000)

$$\sum_{i=1}^{I} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{u}_i) = 0, \ \boldsymbol{D}_i = \frac{\partial \boldsymbol{u}_i}{\partial \boldsymbol{\beta}}, \tag{2}$$

where $\boldsymbol{D}_i$ is the $J\times T$ matrix of partial derivatives of the mean with respect to the regression parameters, T is the number of years, $\boldsymbol{u}_i$ is the predicted crash count at the ith ramp, I is the number of ramp, i indicates the ith ramp, $\boldsymbol{Y}_i$ is the observed crash frequency at the ith ramp, and $\boldsymbol{V}_i$ is the covariance matrix defined as

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\lambda}) \boldsymbol{A}_i^{1/2}, \tag{3}$$

where $\boldsymbol{A}_i$ is a $T\times T$ diagonal matrix with $V(\mu_{it})$ as the tth diagonal element, $\boldsymbol{R}_i(\boldsymbol{\lambda})$ is the $T\times T$ matrix presenting the temporal correlation in repeated observations, and $\boldsymbol{\lambda}$ is the type of correlation with $\boldsymbol{\lambda}=[\lambda_1, \lambda_2, \ldots, \lambda_{n-1}]$ and $\lambda_i=\mathrm{cor}(Y_t, Y_k)$ for t, k=1, 2, …, $n-1$, $t\neq k$.

To solve the model with GEE correctly, every element of the correlation matrix $\boldsymbol{R}_i$ has to be known. However, in many instances, it is not possible to know the proper correlation type for the crash counts per year. To overcome this drawback, Liang and Zeger (1986) proposed a "working" matrix as the correlation matrix to estimate the coefficients. The commonly used correlation structure in the GEE procedure is briefly described as follows:

1. Independent: the independent correlation structure assumes that repeated observations (crash counts in different years) for an exit ramp are independent. In this case, the GEE estimates are the same as the regular GLM in the coefficients but different in the standard errors (SEs).

2. Exchangeable: the exchangeable working correlation assumes constant correlations between any two observations within an exit ramp.

3. Autoregressive: the autoregressive correlation structure weighs the correlation between two observations by their separated time-gap (order of measure). As the gap distance increases, the correlation decreases.

4. Unstructured: the unstructured correlation structure assumes a different correlation between any two observations taken at the same location.

3.3 Model assessment

Traditional goodness-of-fit tests for basic GLM are not valid for the GLM with GEE procedure (Wang and Abdel-Aty, 2006). The cumulative residual tests are conducted to graphically and numerically examine how well the link function fits the dataset. The cumulative residual method has an advantage of being independent on the number of observations as are many other traditional statistical procedures (Hauer, 2004; Wang and Abdel-Aty, 2006; 2008). If the model is correct, the residuals should be centered at zero and the plot of the residuals against any coordinate should exhibit no systematic tendency. The maximum absolute value of the observed cumulative sum and the *P*-value for a Kolmogorov-type supremum test are calculated. A small maximum absolute value and a large *P*-value indicate a better model performance.

The type III analysis has been used to identify a variable's relative significance (Wang and Abdel-Aty, 2006; 2008). The type III χ^2 value for a particular variable is the difference between the generalized score statistic or likelihood ratio statistic for the model with all the variables included and that with this variable excluded. A small *P*-value indicates that the effect of this variable is highly significant.

4 Results

The crash data in this study are shown to be over-dispersed so that the GLM with a NB distribution in the random component is considered to fit the data. Two traditional GLMs based on yearly aggregated and disaggregated crash data are evaluated and a GLM with GEE procedure is fitted. The model estimates are compared and the results are discussed.

4.1 GLM model estimates

Two GLMs are estimated in this section: the first model (GLM 1) uses the average crash count per year across the four years as the dependent variable; and the second model (GLM 2) uses the crash count in each year. The GLMs take the model forms in Eq. (1) and the explanatory variables are carefully selected to determine the final model specifications. The estimates of the two GLMs are shown in Table 3. Only the variables that are significant in at least one model are included.

Table 3 Model estimating results of GLMs

Variable	GLM 1			GLM 2		
	Coefficient	SE	*P*-value	Coefficient	SE	*P*-value
Intercept	0.888	0.345	0.010	−1.992	1.271	0.117
Grade	0.142	0.125	0.259	0.323	0.144	0.025
Logarithm of AADT on mainline	0.207	0.188	0.269	0.520	0.165	0.002
Logarithm of AADT on ramp	−0.099	0.074	0.182	0.244	0.073	0.001
Bad weather ratio	4.007	1.102	<0.0001	3.078	0.456	<0.0001
Right shoulder width	−4.603	2.788	0.099	−0.112	0.062	0.073
Summary statistics						
Level of dispersion, α	0.051			0.1905		
Deviance/degrees of freedom	7.5376			1.0834		
Pearson χ^2/degrees of freedom	7.8978			1.0026		
Akaike information criterion	1565.4303			1003.3095		
Bayesian information criterion	1582.5425			1023.2737		

In the GLM 2, five variables are significantly related to the crash count per year at a 90% confidence level. These variables include the AADT on mainline, AADT on exit ramp, presence of grade, bad weather ratio, and right shoulder width. However, in the GLM 1, only two variables, the bad weather ratio and right shoulder width, are estimated to be significant at a 90% confidence level. The other variables such as AADTs on mainline and ramp are not statistically significant, which is contrary to the intuition. The performance of the GLM 1 would attribute to the impact of a small sample size in the dataset. This is supported by the fact that the GLM 2 has a better statistical fitness than the GLM 1 as shown in Table 3. These results show how the small sample size issue impacts the model estimates and leads to poor model performances.

Traditional GLMs assume that the response variable (crash count in this study) is independent of each other, which may not be true for the longitudinal data with repeated observations over time at each location. Crashes in different years could be intercorrelated due to the unobserved or unconsidered effects of factors associated with a specific exit ramp that did not change over the years. Fig. 2 shows that the correlation exists between crashes of different years. The traditional GLM with yearly disaggregated data developed above did not account for the temporal correlation in the dataset and could result in biased model estimates.

4.2 Model estimates with GEE

The GLM model with GEE procedure is fitted using the yearly disaggregated data to account for the temporal correlation. Four types of correlation structure, which are the independent, exchangeable, autoregressive, and unstructured structure, are explored in the GEE procedure. The estimating results of these models are shown in Table 4. It can be identified that the coefficients and SEs for explanatory variables are consistent between models with different correlation structures. It indicates that the GEE approach has a robust performance and that the estimates would be correct even when the covariance matrix is specified incorrectly (Lord and Persaud, 2000). Though the estimates are similar, the four models produce unequal estimating results, which shows the impacts of different correlation structures in the GEE procedure.

The estimated correlation matrix with a dimension of four for each type of correlation structure is shown in Table 5. The assessments of models with different correlation structures are performed using

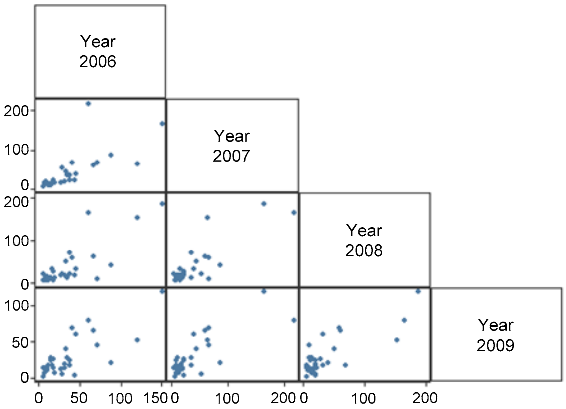

Fig. 2 Correlation of crash counts between years

Table 4 Model estimates of GLM with GEE procedure

Variable	Independent		Exchangeable		Autoregressive		Unstructured	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Intercept	−1.992	2.000	−1.667	2.259	−1.526	2.236	−1.862	2.309
Grade	0.323	0.189	0.377	0.207	0.402	0.209	0.445	0.182
Logarithm of AADT on mainline	0.520	0.229	0.470	0.251	0.465	0.252	0.530	0.250
Logarithm of AADT on ramp	0.244	0.103	0.266	0.110	0.220	0.108	0.181	0.118
Bad weather ratio	3.078	0.496	2.795	0.435	2.863	0.457	2.982	0.423
Right shoulder width	−0.112	0.072	−0.116	0.078	−0.098	0.076	−0.103	0.080
Summary statistics								
Cluster size	4		4		4		4	
Maximum absolute value	23.15		22.24		44.46		59.63	
P-value	0.470		0.618		0.092		0.026	

the cumulative residual test, and the results are shown in Fig. 3. The observed cumulative residuals for working correlation structures are represented by the heavy lines, and the simulated curves are represented by the light lines. The residuals for the GEE with exchangeable correlation structures are centered at zero and the plot of the residuals against any coordinate exhibits no systematic tendency. Also as shown in Table 4, the GEE with exchangeable structure has the smallest maximum absolute value and the largest *P*-value among all the structures. These assessments indicate that the exchangeable structure in the GEE is fairly appropriate to fit the inherent feature of data in this study.

Table 5 Estimated working correlation structures

Correlation structure	Year	2006	2007	2008	2009
Independent	2006	1.000			
	2007	0.000	1.000		
	2008	0.000	0.000	1.000	
	2009	0.000	0.000	0.000	1.000
Exchangeable	2006	1.000			
	2007	0.271	1.000		
	2008	0.271	0.271	1.000	
	2009	0.271	0.271	0.271	1.000
Autoregressive	2006	1.000			
	2007	0.338	1.000		
	2008	0.114	0.338	1.000	
	2009	0.039	0.114	0.338	1.000
Unstructured	2006	1.000			
	2007	0.717	1.000		
	2008	0.455	0.614	1.000	
	2009	0.151	−0.088	−0.016	1.000

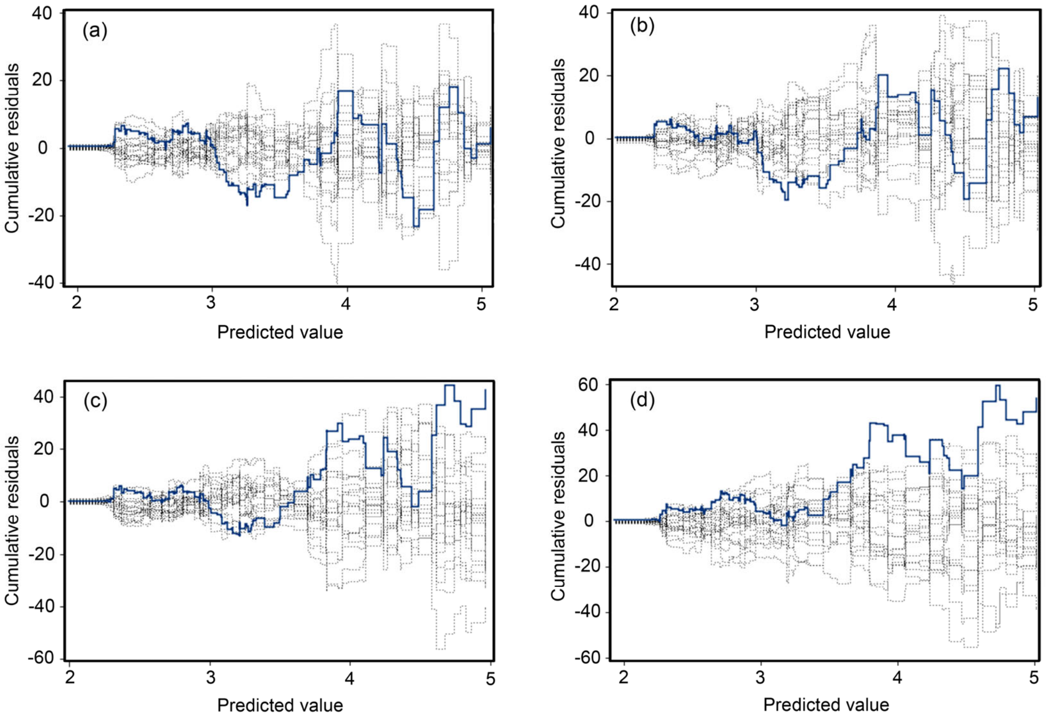

Fig. 3 Model assessments for GEEs with different correlation structures
(a) Independent; (b) Exchangeable; (c) Autoregressive; (d) Unstructured

The exchangeable structure assumes that the correlations between multiple observations are constant. As shown in Table 5, the correlation between two successive observations is 0.271, indicating that there is a significant temporal correlation between crash counts at an exit ramp in different years. The relatively high correlation should not be neglected during the crash modeling procedure. In some previous studies, the autoregressive structure in GEE was found to have the best goodness-of-fit since it assumed that the correlation between observations would decrease as the time-gap increases (Wang and Abdel-Aty, 2006). The different findings on the performance of correlation structure between this study and previous ones would be explained by the different characteristics of crash data that have been used. In this study, it is identified that the exchangeable correlation structure is more consistent than the correlation plots shown in Fig. 2.

In sum, with the results obtained above, it is reasonable to conclude that: (1) there is obvious temporal correlation in the crash data with yearly disaggregated observations; (2) the exchangeable working correlation structure is the most fitted one in the GEE procedure for analyzing the four-year frame data in this study.

4.3 Comparison between models

A comparison on the coefficient and SE of each variable between the GLM 2 and the GLM with GEE shows that though the coefficients are shown to be similar in the two models, the SEs in the GLM with GEE are obviously larger than those in the traditional GLM 2. This result suggests that the temporal correlation contributes to a large amount of SEs for explanatory variables. The increase of SE would decrease the significant level of a variable. In other words, some factors may become insignificant after considering the temporal correlation in the data.

Recall that the temporal correlation in the crash data is generally generated by the unobserved or unconsidered effects of factors that do not change over years on an exit ramp. If the temporal correlation is not properly considered in the modeling procedure, the variation of crash counts could be incorrectly attributable to the variation of observed variables, other than these unobserved effects. In the traditional GLM, the estimated effects of explanatory variables potentially contain some effects of unobserved factors. In this situation, the inferences on the impacts of contributing variables on crashes could still be biased and misleading.

The type III analyses are performed to examine the relative significance of explanatory variables. As shown in Table 6, the type III χ^2 values in the GLM with GEE are generally smaller than that in the GLM 2 and the *P*-values for variables are larger in the GLM with GEE. It indicates that the traditional GLM without accounting for the temporal correlation would overestimate the significance of predicting factors, which is consistent with previous studies (Lord and Persaud, 2000; Wang and Abdel-Aty, 2006; 2008).

As shown in Table 6, the right shoulder width is estimated to be significantly related to crash counts at a 90% confidence level in the GLM 2. However, after considering the temporal correlation in the GLM with GEE, this variable becomes insignificant at the same confidence level. Though more crashes are reported at exit ramps with narrower right shoulders, the large number of crashes would be due to some effects of unobserved factors such as poor pavement or unsafe geometric designs (which are reflected in the temporal correlation) other than the effect of the right shoulder width. If the shoulder width was incorrectly considered to predict the normal safety level for exit ramps, some true hotspots with higher-than-normal crashes could not be identified correctly. Considering

Table 6 Type III analyses for different models

Variable	GLM 2		GLM with GEE (exchangeable)	
	Type III χ^2	*P*-value	Type III χ^2	*P*-value
Curve	4.97	0.0258	2.93	0.0868
Logarithm of AADT on mainline	9.44	0.0021	3.04	0.0814
Logarithm of AADT on ramp	10.30	0.0013	5.91	0.0151
Bad weather ratio	40.41	<0.0001	10.39	0.0013
Right shoulder width	3.22	0.0726	1.99	0.1579

the temporal correlation using the GEE procedure could result in more accurate inferences. It could help safety researchers or agencies make correct decisions to implement countermeasures on dangerous ramps.

4.4 Interpretation of coefficients

The AADTs on mainline and ramps are estimated to be positively related to crash counts at exit ramps. The increase of traffic volume results in an increase of traffic crashes. The coefficient for AADT on ramp is larger than that for AADT on mainline suggesting that a unit increase in traffic volume on an exit ramp could generate more crashes as compared to that on a mainline. The presence of grade will increase the crash counts since the estimated coefficient for the variable is positive. More crashes are likely to occur under bad weather conditions.

Several insignificant explanatory variables were reported to be significant predictors in some studies. For example, the length of the deceleration lane and the length of the exit ramp have been identified to be significantly related to crash counts at freeway exit ramp areas (Chen *et al.*, 2009; 2011; Liu *et al.*, 2010). It could be difficult to tell if the insignificances of these variables in this study reflect the actual situation on the freeways in China or are generated due to the limitation of sample size used for model development. These variables at a ramp do not vary over years so that the disaggregation of data per year could not improve the estimates for these variables. Data with larger sample size are always desirable to obtain more accurate estimates on the relationships between these variables and crash counts.

5 Conclusions and discussion

This study evaluated the application of the GEE to account for the temporal correlation in the crash frequency data. Using four-year crash data at exit ramps on the Guangshen freeway, China, the GLM with GEE was estimated based on yearly disaggregated crash data. For comparison purposes, traditional GLMs were also estimated based on the same dataset.

The results showed that there were significant temporal correlations in the yearly disaggregated crash data used in this study. The GEE procedure captured the correlation among crash counts in different years. The exchangeable correlation structure fitted the data properly. A comparison between the GLM and the GLM with GEE showed that the traditional GLM could underestimate the SEs of explanatory variables and make incorrect inferences on the significance of the variables. The GLM with GEE captured the features of temporal correlation in the data and led to more accurate estimates on the impacts of predictors. In the modeling results, the right shoulder width was identified to be a significant factor in the traditional GLM, but became insignificant after accounting for the temporal correlation in the GLM with GEE. Other contributing factors on crashes at freeway exit ramps included the AADT on mainline, AADT on ramps, presence of grade, and bad weather ratio.

The findings of this study suggest that the GEE is an appropriate approach for modeling crash frequency data with temporal correlation. This approach makes it relatively easy to develop proper and accurate crash prediction models even if the type of temporal correlation is unknown. The GEE procedure also has an advantage that many statistical software packages already have a built-in GEE functionality.

Even though this study showed that using disaggregated crash data results in better model predictions than using aggregated crash data, such a conclusion does not hold true in many situations. This study simply showed that the models with enlarged sample size (by extending data to more than one year) perform better than the models with small sample size. However, it does not mean that the models based on monthly or weekly crash data will definitely outperform those with yearly crash data, because the predicted values of the models are rather different. Detailed experiments and modeling are required to compare the performances of different models based on different temporal segmentations of crash data. Besides, it should be explained that extending the crash data to smaller aggregation intervals may lead to an increase in the number of sections with zero counts, leading possibly to the need for zero-inflated models, since excessive zero counts do not fit the regular Poisson or negative binomial models.

The observations from the same year may be correlated due to unobserved within-year effects, which are termed as the spatial correlation. Though the GEE procedure can successfully account for the temporal correlation in the crash frequency data, it cannot address the spatial correlation that could also exist in the crash data. Recently, researchers have proposed more sophisticated models which can account for the spatial correlation across locations, such as the random-effects model (Shankar *et al.*, 1998; Miaou and Lord, 2003; Quddus, 2008) and the hierarchica/multilevel model (Jones and Jørgensen, 2003; Kim *et al.*, 2007). Considering the spatial correlation in the modeling procedure could improve the model predictions. The authors recommend that future studies could focus on these issues.

References

Bared, J., Giering, G.L., Warren, D.L., 1999. Safety evaluation of acceleration and deceleration lane lengths. *Institute of Transportation Engineers Journal*, **69**(5):50-54.

Bauer, K.M., Harwood, D.W., 1998. Statistical Models of Accidents on Interchange Ramps and Speed-change Lanes. National Technical Information Service, Alexandria, USA, p.1-163.

Chen, H., Liu, P., Lu, J.J., *et al.*, 2009. Evaluating the safety impacts of the number and arrangement of lanes on freeway exit ramps. *Accident Analysis and Prevention*, **41**(3): 543-551. [doi:10.1016/j.aap.2009.01.016]

Chen, H., Zhou, H., Zhao, J., *et al.*, 2011. Safety performance evaluation of left-side off-ramps at freeway diverge areas. *Accident Analysis and Prevention*, **43**(3):605-612. [doi:10.1016/j.aap.2010.08.019]

El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. *Accident Analysis and Prevention*, **41**(4):820-828. [doi:10.1016/j.aap.2009.04.005]

Hauer, E., 1997. Observational Before-after Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Pergamon Press, Elsevier Science Ltd., Netherlands.

Hauer, E., 2004. Statistical road safety modeling. *Transportation Research Record: Journal of the Transportation Research Board*, **1897**(1):81-87. [doi:10.3141/1897-11]

Hu, G., Wen, M., Baker, T.D., *et al.*, 2008. Road-traffic death in China, 1985-2005: threat and opportunity. *Injury Prevention*, **14**(2):129-130. [doi:10.1136/ip.2007.017822]

Jin, S., Huang, Z., Tao, P., *et al.*, 2011. Car-following theory of steady-state traffic flow using time-to-collision. *Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)*, **12**(8):645-654. [doi:10.1631/jzus.A1000518]

Jin, S., Wang, D., Ma, D., 2013. Using LBG quantization for particle-based collision detection algorithm. *Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)*, **14**(4):231-243. [doi:10.1631/jzus.A1200218]

Jones, A.P., Jørgensen, S.H., 2003. The use of multilevel models for the prediction of road accident outcomes. *Accident Analysis and Prevention*, **35**(1):59-69. [doi:10.1016/S0001-4575(01)00086-0]

Kim, D.G., Lee, Y., Washington, S., *et al.*, 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. *Accident Analysis and Prevention*, **39**(1):125-134. [doi:10.1016/j.aap.2006.06.011]

Li, Z., Liu, P., Wang, W., *et al.*, 2012. Using support vector machine models for crash injury severity analysis. *Accident Analysis and Prevention*, **45**:478-486. [doi:10.1016/j.aap.2011.08.016]

Li, Z., Chung, K., Cassidy, M., 2013. Collisions in freeway traffic: the influence of downstream queues and interim means to address it. *Transportation Research Record: Journal of the Transportation Research Board*, **2396**(1): 1-9. [doi:10.3141/2396-01]

Li, Z., Ahn, S., Chung, K., *et al.*, 2014. Surrogate safety measure for evaluating rear-end collision risk related to kinematic waves near freeway recurrent bottlenecks. *Accident Analysis & Prevention*, **64**:52-61. [doi:10.1016/j.aap.2013.11.003]

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1):13-22. [doi:10.1093/biomet/73.1.13]

Liu, P., Chen, H., Lu, J., *et al.*, 2010. How arrangement of lanes on freeway mainlines and ramps affects safety of freeways with closely spaced entrance and exit ramps. *Journal of Transportation Engineering*, **136**(7):614-622. [doi:10.1061/(ASCE)TE.1943-5436.0000127]

Lord, D., Persaud, B.N., 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. *Transportation Research Record: Journal of the Transportation Research Board*, **1717**(1): 102-108. [doi:10.3141/1717-13]

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, **44**(5):291-305. [doi:10.1016/j.tra.2010.02.001]

Luoma, J., Sivak, M., 2012. Road-safety Management in Brazil, Russia, India, and China. University of Michigan Transportation Research Institute, Michigan, USA, UMTRI-2012-1.

Ma, J.M., Kockelman, K.M., 2006. Bayesian multivariate Poisson regression for models of injury count, by severity. *Transportation Research Record: Journal of the Transportation Research Board*, **1950**(1):24-34. [doi:10.3141/1950-04]

Ma, S., Li, Q., Zhou, M., *et al.*, 2012. Road traffic injury in China: a review of national data sources. *Traffic Injury Prevention*, **13**(sup1):57-63. [doi:10.1080/15389588.2011.633945]

McCartt, A.T., Northrup, V.S., Retting, R.A., 2004. Types and characteristics of ramp-related motor vehicle crashes on urban interstate roadways in northern Virginia. *Journal of*

Safety Research, **35**(1):107-114. [doi:10.1016/j.jsr.2003.09.019]

Miaou, S.P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes. *Transportation Research Record: Journal of the Transportation Research Board*, **1840**(1):31-40. [doi:10.3141/1840-04]

Park, E., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record: Journal of the Transportation Research Board*, **2019**(1):1-6. [doi:10.3141/2019-01]

Quddus, M.A., 2008. Time series count data models: an empirical application to traffic accidents. *Accident Analysis and Prevention*, **40**(5):1732-1741. [doi:10.1016/j.aap.2008.06.011]

Saenghaengtham, N., Kanongchaiyos, P., 2006. Using LBG quantization for particle-based collision detection algorithm. *Journal of Zhejiang University-SCIENCE A*, 7(7):1225-1232. [doi:10.1631/jzus.2006.A1225]

Shankar, V.N., Albin, R.B., Milton, J.C., *et al.*, 1998. Evaluating median cross-over likelihoods with clustered accident counts: an empirical inquiry using random effects negative binomial model. *Transportation Research Record: Journal of the Transportation Research Board*, **1635**(1):44-48. [doi:10.3141/1635-06]

Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. *Accident Analysis and Prevention*, **38**(6):1137-1150. [doi:10.1016/j.aap.2006.04.022]

Wang, X., Abdel-Aty, M., 2008. Modeling left-turn crash occurrence at signalized intersections by conflicting patterns. *Accident Analysis and Prevention*, **40**(1):76-88. [doi:10.1016/j.aap.2007.04.006]

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2010. Statistical and Econometric Methods for Transportation Data Analysis. Chapman & Hall/CRC, Boca Raton, USA.

中文概要：

本文题目： 基于广义估计方程的时间相关性事故频次建模

Application of generalized estimating equations for crash frequency modeling with temporal correlation

研究目的： 采用广义估计方程模型对存在时间相关性的事故频次数据进行建模，并与传统广义线性模型的估计效果进行对比。

创新要点： 通过广义估计方程来考虑事故频次建模中数据的时间相关性，从而提高参数估计准确度以及模型预测精度。

研究方法： 基于 4 年高速公路交通事故频次数据，建立考虑时间相关性的广义估计方程以及传统的广义线性模型，并采用统计指标对模型效果进行对比。

重要结论： 1. 事故频次数据样本量对预测精度影响很大；2. 广义估计方程能够有效考虑事故频次数据中存在的时间相关性；3. 广义估计方程的参数估计比传统广义线性模型更准确，且精度更高。

关键词组： 广义估计方程；事故频次；时间相关；广义线型模型