Journal of Southeast University (English Edition) Vol. 30, No. 1, pp. 96 - 100 Mar. 2014 ISSN 1003—7985

Conditional autoregressive negative binomial model for analysis of crash count using Bayesian methods

Xu Jian[1,2] Sun Lu[1,3]

([1]School of Transportation, Southeast University, Nanjing 210096, China)
([2]Center for Transportation Research, University of Texas at Austin, Austin 78712, USA)
([3]Department of Civil Engineering, Catholic University of America, Washington DC 20064, USA)

Abstract: In order to improve crash occurrence models to account for the influence of various contributing factors, a conditional autoregressive negative binomial (CAR-NB) model is employed to allow for overdispersion (tackled by the NB component), unobserved heterogeneity and spatial autocorrelation (captured by the CAR process), using Markov chain Monte Carlo methods and the Gibbs sampler. Statistical tests suggest that the CAR-NB model is preferred over the CAR-Poisson, NB, zero-inflated Poisson, zero-inflated NB models, due to its lower prediction errors and more robust parameter inference. The study results show that crash frequency and fatalities are positively associated with the number of lanes, curve length, annual average daily traffic (AADT) per lane, as well as rainfall. Speed limit and the distances to the nearest hospitals have negative associations with segment-based crash counts but positive associations with fatality counts, presumably as a result of worsened collision impacts at higher speed and time loss during transporting crash victims.

Key words: traffic safety; crash count; conditional autoregressive negative binomial model; Bayesian analysis; Markov chain Monte Carlo

doi: 10.3969/j.issn.1003-7985.2014.01.018

With the increase in the number of vehicles, it is interesting and commendable that currently fatalities are decreasing every year in China, the reason of which can be attributed to the optimization of roadway designs, more safety vehicles, as well as many researches of crashes and the contributing factors. However, still 210 812 reported crashes and 62 387 reported fatalities occurred on roadways in 2011 in China according to official reports[1], demanding the further improvement of transportation safety to reduce the traffic accidents and fatalities.

The possible access to understand the elements of crashes is to develop statistical analysis methods used to distinguish the significant factors, which can be utilized to provide an optimality criterion to policy makers. During the past several years, numerous methods for analyzing crash counts were proposed[2-6]. The earliest approach for crash count data is the Poisson model[7], and then it gives rise to more flexible alternatives, e.g., the negative binomial (NB) model[8], the GIS-based Bayesian approach[9], the finite mixture regression model[10], and the quantile regression method[11]. Most of the regression methods applied to model crash counts, however, are focused on aspatial (i.e. non-spatial) analysis. Applied work in aspatial models may not be able to capture spatial heterogeneity and spatial dependence at neighborhood areas, a frequently happening issue in crash counts. This leads to the development of alternative methodologies that focus on spatial modeling in the past few decades. Early pioneering work on spatial modeling is reported by Besag[12], and is further enriched by LeSage et al[13-16]. Anselin[17] provided two specifications of spatial models, spatial error model (SEM) (i.e., the spatial autocorrelation model (SAC)) and the spatial lag model (SLM) (i.e., the spatial autoregressive model (SAR)) that is a special type of conditional autoregressive (CAR) model, at least in a continuous-response setting.

The primary objective of this study is to develop associations between crash counts on homogeneous segments and the contributing factors, using a negative binomial (NB)-based conditional autoregressive model (CAR) which allows for overdispersion, unobserved heterogeneity and spatial autocorrelation. The Bayesian estimation is employed, using Markov chain Monte Carlo methods and the Gibbs sampler. The independent variables consist of traffic characteristics, roadway design and built environments, and the data are derived from on-system highways of Austin, TX, USA in the year 2010. Meanwhile, the exposure variable and the dummy variable are also considered.

1 Model Structure

As described before, there are two specifications of spatial models: the spatial autocorrelation model and the spatial autoregressive model. The general formulation of

Received 2013-08-23.

Biographies: Xu Jian (1985—), male, graduate; Sun Lu (corresponding author), male, doctor, professor, sunl@cua.edu.

Foundation items: The National Science Foundation by Changjiang Scholarship of Ministry of Education of China (No. BCS-0527508), the Joint Research Fund for Overseas Natural Science of China (No. 51250110075), the Natural Science Foundation of Jiangsu Province (No. SBK200910046), the Postdoctoral Science Foundation of Jiangsu Province (No. 0901005C).

Citation: Xu Jian, Sun Lu. Conditional autoregressive negative binomial model for analysis of crash count using Bayesian methods[J]. Journal of Southeast University (English Edition), 2014, 30(1): 96 - 100. [doi: 10.3969/j.issn.1003-7985.2014.01.018]

the spatial autoregressive model for cross-sectional spatial data is

$$\boldsymbol{y}_i = \rho \boldsymbol{W}_1 \boldsymbol{y}_i + \boldsymbol{x}_i \beta + \boldsymbol{\phi} \tag{1}$$

where $\boldsymbol{y}_i$ contains an $n \times 1$ vector of dependent variables; ρ is the spatial lag coefficient; $\boldsymbol{W}_1$ is the spatial weights matrix; $\boldsymbol{\phi}$ is the error term for spatial dependence; $\boldsymbol{x}_i$ represents the matrix of independent variables.

$$\boldsymbol{\phi} = \lambda \boldsymbol{W}_2 \boldsymbol{\phi}_i + \boldsymbol{\varepsilon} \tag{2}$$

where λ is the spatial autoregressive coefficient; $\boldsymbol{W}_2$ is a known spatial weights matrix like $\boldsymbol{W}_1$, usually containing the first-order contiguity relationships; $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$. The SAR model tends to be difficult to develop for limited-response frameworks, especially when dealing with large scale problems involving a large amount of observations, and yields parameter estimates similar to those estimated from the CAR model. Moreover, due to faster computation, the CAR model is preferred in spatial analysis over the SAR model. Under the MRF assumption, the conditional probability density function of the univariate CAR model is [18]

$$f(\psi_i \mid \psi_{i^*}) = \sqrt{\frac{1}{2\pi\sigma_i^2}} \exp\left\{ -(2\sigma_i^2)^{-1} \left[(\psi_i - \mu_i) - \alpha \sum_{i^* \in N_i} w_{ii^*} (\psi_{i^*} - \mu_{i^*}) \right]^2 \right\} \tag{3}$$

The joint probability density function is

$$f(\boldsymbol{\psi}) = \frac{1}{(2\pi)^{n/2} \det[(\boldsymbol{I} - \alpha \boldsymbol{W})^{-1} \boldsymbol{D}_{\sigma^2}]^{1/2}} \cdot \exp\left[-\frac{1}{2} (\boldsymbol{\psi} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{D}_{\sigma^2}^{-1} (\boldsymbol{I} - \alpha \boldsymbol{W}) (\boldsymbol{\psi} - \boldsymbol{\mu}) \right] \tag{4}$$

where i^+ denotes the neighboring locations of area i; $i, i^+ = 1, 2, \cdots, n$ represent the areas for analysis; ψ_i indicates the spatially autocorrelated variable (a random variable) and $\boldsymbol{\psi} = \{\psi_1, \psi_2, \cdots, \psi_n\}^{\mathrm{T}}$; μ_i is the expected value (i.e., mean value) of $\boldsymbol{x}_i$ and $\boldsymbol{\mu} = \{\mu_1, \mu_2, \cdots, \mu_n\}^{\mathrm{T}}$; σ_i^2 is the conditional variance and $\boldsymbol{D}_{\sigma^2} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$; w_{ii^*} is the weight that describes the distance or contiguity relationships between i and i^+; $\boldsymbol{W} = \{w_{11^*}, w_{22^*}, \cdots, w_{nn^*}\}^{\mathrm{T}}$, $w_{ii} = 0$ and $w_{ii^*} = w_{i^*i}$; α is a smoothing parameter defined as measuring spatial association. This study relies on the proper CAR specification for crash rates with NB-based count model settings. The CAR model is estimated using Bayesian Markov chain Monte Carlo techniques coded in WinBUGS and employs an adjacency matrix, as shown before. The crash rate is modeled as a function of the covariates:

$$\lambda_i = E_i^{\tau} \exp(\beta_0 + \beta_k X_{ik} + \psi_i + \varepsilon_i) \tag{5}$$

where E_i is the exposure variable, which represents vehicle miles traveled (VMT) in this study; τ denotes an unknown parameter for the exposure measure; β_0 is the intercept term; β_k denotes the coefficient of the k-th covariate; X_{ik} are indicators for the k-th covariate for segment i; ψ_i follows the proper CAR prior, as described before; ε_i is a random error that has a gamma distribution, that is, $\varepsilon_i \sim \Gamma(\theta, \theta)$.

2 Data Description

In this study, roadways and crash data sets of Austin City in USA in 2010 are used to examine the associations between crash counts on mainlanes and the contributing factors. The roadways in this study are on-system highways, containing interstate highways, US highways, state highways, farm-to-market roadways, etc. In order to avoid the modifiable areal unit problem (MAUP) [19], roadways are split into 1 824 homogeneous segments where geometric characteristics are coincident, as shown in Fig. 1. Most segments have a length of 0 to 1.6 km and occupy more than 90% of the whole sample. The average of the segment length on mainlanes is 0.459 km. After merging crashes and segments, 1 413 crashes on mainlanes are matched.

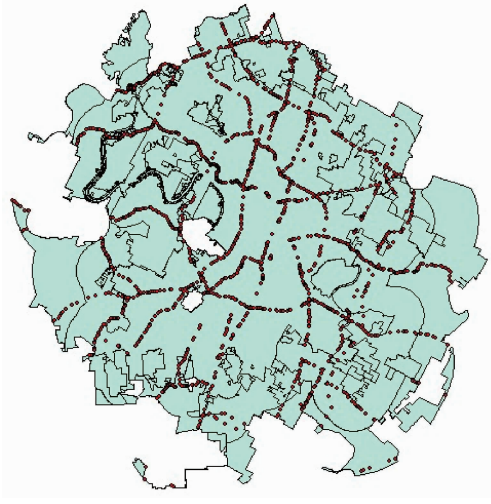

Fig. 1 Distribution of homogeneous segments in Austin (Spots are the center points of segments)

In this study, the dependent variable is the number of crashes, while the exposure variable captures VMT, which is a key crash exposure term (since crash counts closely correlate with VMT, everything else remaining constant), and simply the product of AADT, segment length, and 365 days per year. The dependent variable set contains both continuous and categorical variables, as shown in Tab. 1. The indicator for curvature is a dummy variable, that is, if the answer is yes, it equals 1, and 0 otherwise. In addition, traffic characteristics allow for AADT, speed limit, and the percentage of truck AADT. In the past research, environments, especially distances to the nearest hospitals, were rarely employed for the contributing factors to analyze the associations of crash counts. In this study, hospitals are collected for analysis; meanwhile, the distances of which to segments are computed by ArcGIS, as shown in Fig. 2. The data of annual rainfall obtained from the US Natural Resources Information System are also collected for analysis. It is noted that

it would be best to match the year 2010 crashes to the same year rainfall data, however such information is unavailable, and we cannot find out the data. According to the climate history in Texas, the annual rainfall changed a little, so 1961—1990 average rainfall is used instead. Fig. 3 depicts the distribution of the annual rainfall in Austin.

Tab. 1 Summary statistics of variables for segments

Variable type	Variable name	Mean	Standard deviation	Min	Max
Dependent variable	Crash count	0.775	3.172	0	182
Exposure variable	VMT/(veh · km)	2 566.5	11 668.6	0	895 662
Covariates	Average shoulder width/m	1.723	0.985	0	9.144
	Number of lanes	2.536	1.130	1	13
	Median width/m	1.974	5.523	0	113.4
	Indicator for curvature (1 = yes, 0 = otherwise)	0.418	0.488	0	1
	Curve length/km	0.061	0.136	0	5.194
	Degree of curve/(°)	1.345	3.873	0	72.550
	AADT per lane/(veh · d^{-1})	1 972	2 834	0	34 267
	Percentage of truck AADT/%	8.254	5.982	0	6.422
	Speed limit/(km · h^{-1})	88.67	13.99	8	128
	Rainfall/m	0.872	0.267	0.580	1.194
	Distance to the nearest hospital/km	10.66	7.832	0.310	22.36

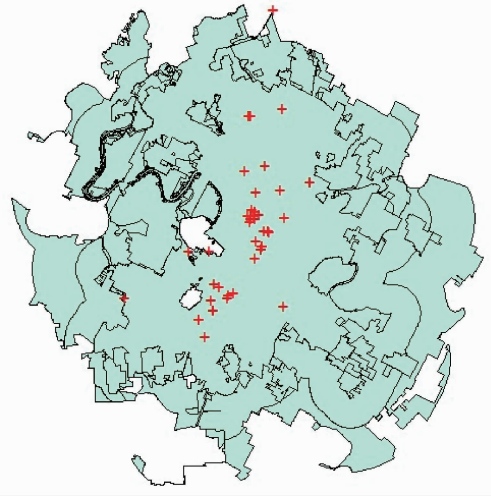

Fig. 2 Distribution of hospitals in Austin

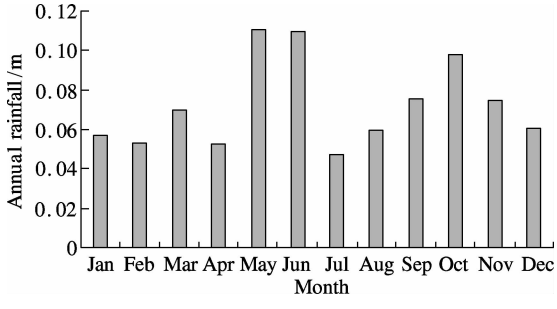

Fig. 3 Distribution of annual rainfall in Austin

3 Estimation Results and Discussion

This section discusses the results of the associations between the contributing factors and the crash counts on mainlanes in Austin. Tab. 2 shows the parameter estimates of the CAR model for crash counts, based on a total number of 5 000 draws in WinBUGS.

The association between crash exposure (VMT) and crash rates is estimated to be nonlinear (average exponent $\tau = 0.658$ for mainlanes), which follows prior expectations. After controlling the exposure variable (VMT), other covariates regarding crash rates are estimated, which can be seen in Tab. 2.

Elasticities for total crash counts and fatal crash counts are computed as the average percentage change in the mean crash rate per 1% change in the k-th variable. As shown in Tab. 2, crash counts are estimated to have a statistically and practically significant spatial autocorrelation coefficient of 0.624 (that is $\alpha = 0.624$). The number of lanes, curve length, AADT per lane, and rainfall have positive impacts on the mean crash rates for mainlanes, while the remaining variables all exhibit negative impacts on the mean crash rates. The elasticity of −0.123 is found to be that of the curve indicator variables, implying that, holding everything else constant at their means, the mean crash rate is estimated to drop by 0.123 when the indicator variable switches from 0 to 1. The result confirms that the roadway curvature has negative effects on crash rates, which is consistent with the findings of some other studies [5-6].

Interestingly, the speed limit on mainlanes exhibits negative mean elasticities, implying that higher speed limits are associated with lower mean crash rates, as found in Ref. [4]. However, the speed limit has a positive effect on fatality rates, as shown in Tab. 2. Rainfall intensity is estimated to be positively associated with crash rates, and an increase of 1% rainfall will result in an increase of 8.622 in crash rates and an increase of 0.283 in fatality rates. As discussed previously, the distances to hospitals rarely appear as contributing factors in the crash modeling literature. It is found that the distances to the nearest hospitals have a negative impact on the mean crash rates, which suggests that shorter distances lead to higher crash rates, however, as expected, positive associations with fatal crash rates (presumably due to more severe collision impacts at higher speeds and time lost in transporting crash victims to an emergency room).

Tab. 2 Estimation results of CAR-NB model for crash and fatal counts

Variable	Mean coefficient	Standard deviation	Pseudo T-statistics	MC error	Elasticity (total)	Elasticity (fatal)
Average shoulder width/m	−0.063	0.021	−0.243	0.003	−0.048	−0.005
Number of lanes	1.535	0.737	5.346	0.003	0.408	0.123
Median width/m	−0.038	0.078	−0.709	0.002	−0.032	−0.007
Indicator for curvature	−0.442	0.218	−2.364	0.002	−0.123	−0.034
Curve length/km	2.672	0.903	4.532	0.003	0.105	0.029
Degree of curve/(°)	−0.162	0.084	−0.677	0.002	−0.018	−0.005
AADT per lane/(veh · d^{-1})	0.007	0.035	0.318	0.000	0.173	0.083
Percentage of truck AADT/%	−0.248	0.138	−0.844	0.003	−0.194	−0.034
Speed limit/(km · h^{-1})	−0.308	0.988	−5.306	0.002	−2.802	0.708
Rainfall/m	0.138	0.027	0.387	0.003	8.622	0.283
Distance to the nearest hospital/km	−0.109	0.052	−0.233	0.002	−0.148	0.072
α	0.624	0.382	0.253	0.003		
τ	0.658	0.134	0.531	0.002		
θ	1.284	0.556	2.408	0.002		
DIC			865.72			

In this study, the CAR-NB model is compared with another spatial model (CAR-Poisson) and some aspatial models (NB, zero-inflated NB and zero-inflated Poisson), as shown in Tab. 3.

Tab. 3 Comparison of results using aspatial models and spatial models

Measures	CAR-NB	CAR-Poisson	NB	Zero-inflated NB	Zero-inflated Poisson
DIC	865.72	1 858.72	1 522.67	1 032.54	1 276.08
Mean LR	−432.63	−1 072.84	−872.85	−693.83	−764.27
Moran's I of residuals	0.012 ($P > \|z\| = 0.044$)	0.393 ($P > \|z\| = 0.017$)	0.225 ($P > \|z\| = 0.308$)	0.152 ($P > \|z\| = 0.036$)	0.27 ($P > \|z\| = 0.075$)

The deviance information criterion (DIC), as a generalization of the Akaike information criterion (AIC), can be used to compare the goodness-of-fit and complexity of different models estimated under a Bayesian framework. The DIC equation is

$$\mathrm{DIC} = D(\bar{\theta}) + 2p_D = \bar{D} + p_D$$

where $D(\bar{\theta})$ is the deviance evaluated at $\bar{\theta}$ which is the posterior mean of the parameters; p_D is the effective number of parameters in the model; $\bar{D}$ is the posterior mean of the deviance statistic $D(\theta)$. With regards to the model superiority and complexity, the lower the DIC, the better the model[20]. Tab. 3 also presents the log likelihood values, which are used in the likelihood ratio chi-square to test whether all predictors' regression coefficients in the model are simultaneously zero. Meanwhile, Moran's I is also considered, which is a measure of spatial autocorrelation developed by Moran[21]. Negative (positive) values indicate negative (positive) spatial autocorrelation and the values range from −1 (indicating perfect dispersion) to +1 (perfect correlation).

It is observed that the CAR-NB model has the lowest DIC and Moran's I of residuals among these tested models. Meanwhile, mean log likelihood values of the CAR-NB model are the largest. The statistical tests suggest that the CAR-NB model is preferred over the CAR-Poisson, NB, zero-inflated Poisson, zero-inflated NB models due to its lower prediction errors and more robust parameter inference. It can be found that the negative binomial models in Tab. 3 are better than the Poisson models due to the fact that overdispersion actually exists in the data.

4 Conclusions

1) Statistical tests of DIC, log likelihood and Moran's I suggest that the CAR-NB model is preferred over the CAR-Poisson, NB, zero-inflated Poisson, zero-inflated NB models, while the negative binomial models are better than the Poisson models.

2) The association between crash exposure (VMT) and crash rates is estimated to be nonlinear (average exponent $\tau = 0.658$ for mainlanes), with crash rates effectively falling as VMT rises.

3) The number of lanes, curve length, AADT per lane, and rainfall have positive impacts on crash count, while the remaining variables all exhibit negative impacts.

4) The distances to the nearest hospitals and the speed limit have negative associations with segment-based crash counts but positive associations with fatality counts, presumably as a result of time loss during transporting crash victims and worsened collision impacts at higher speeds.

References

[1] Traffic Management Bureau of the Ministry of Public Security of the People's Republic of China. Road traffic accident statistics annual report of the People's Republic of China (2010) [R]. Wuxi: Traffic Management Research Institute of the Ministry of Public Security, 2011. (in Chinese)

[2] Qu X, Guo T, Wang W, et al. Measuring speed consistency for freeway diverge areas using factor analysis [J]. *Journal of Central South University: Science and Technology*, 2013, **20**(1): 837−840. (in Chinese)

[3] Pei Y L, Ma J. Research on countermeasures for road condition causes of traffic accidents [J]. *China Journal of Highway and Transport*, 2003, **16**(4): 77 - 82.
[4] Ma J, Kockelman K M, Damien P. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods [J]. *Accident Analysis and Prevention*, 2008, **40**(3): 964 - 975.
[5] Quddus M A, Wang C, Ison S G. Road traffic congestion and crash severity: econometric analysis using ordered response models [J]. *Journal of Transportation Engineering*, 2010, **136**(5): 424 - 435.
[6] Wang C, Quddus M A, Ison S G. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model [J]. *Accident Analysis and Prevention*, 2011, **43**(6): 1979 - 1990.
[7] Jovanis P, Chang H L. Modeling the relationship of accidents to miles traveled [J]. *Transportation Research Record*, 1986, **1068**: 42 - 51.
[8] Lord D. The prediction of accidents on digital networks: characteristics and issues related to the application of accident prediction models [D]. Toronto: University of Toronto, 2000.
[9] Li L, Zhu L, Daniel Z S. A GIS-based Bayesian approach for analyzing spatial-temporal patterns of intra-city motor vehicle crashes [J]. *Journal of Transport Geography*, 2007, **15**(4): 274 - 285.
[10] Park B J, Lord D. Application of finite mixture models for vehicle crash data analysis [J]. *Accident Analysis and Prevention*, 2009, **41**(4): 683 - 91.
[11] Qin X, Reyes P. Conditional quantile analysis for crash count data [J]. *Journal of Transportation Engineering*, 2011, **137**(9): 601 - 607.
[12] Besag J E. Nearest-neighbour systems and the auto-logistic model for binary data [J]. *Journal of the Royal Statistical Society, Series B: Methodological*, 1972, **34**(1): 75 - 83.
[13] LeSage J P. Spatial econometrics [EB/OL]. (1999) [2013-03-15]. http://www.spatial-econometrics.com/.
[14] Miaou S, Song J J, Malick B. Roadway traffic crash mapping: a space-time modeling approach [J]. *Journal of Transportation and Statistics*, 2003, **6**(1): 33 - 57.
[15] Quddus M A. Modeling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data [J]. *Accident Analysis and Prevention*, 2008, **40**(4): 1486 - 1497.
[16] Wang Y, Kockelman K M. A conditional-autoregressive count model for pedestrian crashes across neighborhoods [C/CD]//*The 92nd Annual Meeting of the Transportation Research Board*. Washington DC, USA, 2013.
[17] Anselin L. *Spatial econometrics: methods and models* [M]. Dordrecht: Kluwer Academic Publishers, 1988.
[18] Mariella L, Tarantino M. Spatial temporal conditional auto-regressive model: a new autoregressive matrix [J]. *Australian Journal of Statistics*, 2010, **39**(3): 223 - 244.
[19] Openshaw S. The modifiable areal unit problem [J]. *Concepts and Techniques in Modern Geography*, 1983, **38**: 39 - 41.
[20] Spregelhalter D J, Best N G, Carlin B P, et al. Bayesian measures of model complexity and fit [J]. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 2002, **64**(4): 583 - 639.
[21] Moran P A P. Notes on continuous stochastic phenomena [J]. *Biometrika*, 1950, **37**(1): 17 - 23.

用于交通事故分析的基于贝叶斯方法的条件自回归负二项模型

徐 建[1,2] 孙 璐[1,3]

([1] 东南大学交通学院,南京 210096)
([2]Center for Transportation Research, University of Texas at Austin, Austin 78712, USA)
([3]Department of Civil Engineering, Catholic University of America, Washington DC 20064, USA)

摘要:为了改进用于分析大量影响因素的交通事故模型,采用基于马尔可夫链蒙特卡罗法和吉布斯抽样的条件自回归负二项模型来拟合过度散布性(由负二项过程拟合)、未观察异质性和空间相关性(由条件自回归过程拟合).统计检验显示,由于具有更小的预测误差和更强的参数估计,条件自回归负二项模型优于条件自回归泊松模型、负二项模型、零膨胀泊松模型和零膨胀负二项模型.研究结果表明,交通事故率和死亡人数与车道数、曲线长度、车道年平均日交通量和降雨量成正比.最大限速和最近医院距离与交通事故次数成反比,而与死亡事故次数成正比,这可能是由于过高的速度会引发更严重的事故以及救援伤者时丧失较长时间.

关键词:交通安全;交通事故数;条件自回归负二项模型;贝叶斯分析;马尔可夫链蒙特卡罗
中图分类号:U491.31