Safety Science 70 (2014) 63–69

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/ssci

Examining signalized intersection crash frequency using multivariate zero-inflated Poisson regression

Chunjiao Dong [a,*], Stephen H. Richards [a,1], David B. Clarke [a,2], Xuemei Zhou [b,3], Zhuanglin Ma [c]

[a] Center for Transportation Research, The University of Tennessee, 600 Henley Street, Knoxville, TN 37996, USA
[b] Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, 4800 Cao An Road 201804, Shanghai, PR China
[c] School of Automobile, Chang'an University, Nan Er Huang Zhong Duan, Xi'an 710064, Shaanxi, PR China

ARTICLE INFO

Article history:
Received 12 November 2013
Received in revised form 30 April 2014
Accepted 9 May 2014
Available online 2 June 2014

Keywords:
Crash frequency
Geometric design
MZIP model
Bayesian method

ABSTRACT

In crash frequency studies, correlated multivariate data are often obtained for each roadway entity longitudinally. The multivariate models would be a potential useful method for analysis, since they can account for the correlation among the specific crash types. However, one issue that arises with this correlated multivariate data is the number of zero counts increases as crash counts have many categories. This paper describes a multivariate zero-inflated Poisson (MZIP) regression model as an alternative methodology for modeling multivariate crash count data by severity. The Bayesian method is employed to estimate the model parameters. Using this Bayesian MZIP model, we can take into account correlations that exist among different severity levels. Our new method also can cope with excess zeros in the data, which is a common phenomenon found in practice. The proposed model is applied to the multivariate crash counts obtained from intersections in Tennessee for five years. The results reveal that, compared to the univariate ZIP models and multivariate Poisson-lognormal (MVPLN) models, the MZIP models provide the best statistic fit and have the smallest estimation bias. Apart from the improvement in goodness of fit, the results of the MZIP models show promise toward the goal of obtaining more accurate estimates by accounting for excess zeros in correlated count data.

1. Introduction

Intersections present a complicated and hazardous roadway environment to drivers. The presence of signals, guide signs for street names, indications of upcoming turn lanes, conflict traffic, exclusive left turn and right turn lanes, and other paraphernalia associated with intersections create a high degree of conflict that leads to higher crash numbers. In addition, intersections with closely spaced decision points, intensive land use, complex design features, and heavy traffic may cause information overload or driver confusion, resulting in an inadequate understanding of the driving situation and subsequently crashes. Based on the Fatality Analysis Reporting System (FARS) and National Automotive Sampling System-General Estimates System (NASS-GES) data, about 40% of the estimated 5,338,000 crashes during 2011 in the United States were intersection-related. Of those intersection crashes, about 36% occurred at signalized intersections. Furthermore, signalized intersections also tended to experience more severe crashes. Injury crashes accounted for 33.2% of reported signalized intersection crashes, compared to 25.2% for non-signalized intersection crashes. Many studies have shown that intersections are among the most dangerous locations of a roadway network. Therefore, there is a need to understand the factors that contribute to crashes at such locations.

However, some intersections might have only fatal crashes and others might have only incapacitating crashes, non-incapacitating crashes, possible injury crashes, or property damage only crashes. The differences might be caused by geometric features and traffic characteristics from intersection to intersection. Therefore, understanding which factor is significant to the specific crash types and what differences existing between crash types becomes a necessary subject needing to be studied. Since injury crashes cause very serious problems, the goal is to reduce the severe crash frequencies, such as fatal crashes and incapacitating crashes, when the total crash numbers are controlled at certain level.

This paper proposes a MZIP regression model to estimate the relationship between intersection geometric features, traffic

* Corresponding author. Tel.: +1 865 974 1826.
E-mail addresses: cdong5@utk.edu (C. Dong), stever@utk.edu (S.H. Richards), dbclarke@utk.edu (D.B. Clarke), zhouxm@tongji.edu.cn (X. Zhou), mazhuanglin@126.com (Z. Ma).
[1] Tel.: +1 865 974 0724.
[2] Tel.: +1 865 974 1813.
[3] Tel.: +86 021 6958 3001.

http://dx.doi.org/10.1016/j.ssci.2014.05.006
0925-7535/

factors, and crash counts across severity. The primary objective is to conduct a new alternative multivariate model to account for excess zeros in correlated count data. The secondary objective is to investigate which factors significantly contribute to the crash counts across severity. The last objective is to examine if there is any difference in the specific types of crashes for the same factors and how to control the factors to reduce severe crash frequencies under certain level crash frequencies.

2. Literature review

To deal with the data and methodological issues associated with crash frequency data, a wide variety of count data models (Zhang et al., 2012; Chang and Chen, 2005) have been applied over the years. Models to estimate crash frequencies on roadway segments or at intersections fall into two broad categories. One category includes conventional univariate regression models, such as the Poisson models, Poisson-gamma (negative binomial) models, Poisson-lognormal models, zero-inflated models, Conway–Maxwell–Poisson models, gamma models, and generalized estimating equation models. The second category includes potentially more realistic specifications such as generalized additive models, random-effects models, negative multinomial models, random-parameters models, bivariate/multivariate models, finite mixture/Markov switching models, duration models, and hierarchical/multilevel models (for a complete review of this literature see Lord and Mannering, 2010).

In crash frequency analysis, modeling the frequencies of specific types of crashes cannot be done with independent count models, since the frequencies of specific crash types are not independent. When one wishes to model specific types of crash counts (for example, the number of crashes resulting in fatalities, injuries, etc.), multivariate models become necessary because they explicitly consider the correlation among the severity levels for each intersection (Miaou and Song, 2005; Bijleveld, 2005; Song et al., 2006). Ma and Kockelman (2006) used a multivariate Poisson (MVP) regression model to estimate the injury count by severity level. Positive correlation in unobserved factors affecting count outcomes was found across severity levels, resulting in a statistically significant assistive latent term. Several researchers (Park and Lord, 2007; Ma et al., 2008; El-Basyouny and Sayed, 2009; Dong et al., 2014) employed a multivariate Poisson-lognormal (MVPLN) approach as an improved model to describe the relationship between factors and crash counts. Anastasopoulos et al. (2012) proposed a multivariate tobit regression model to handle left-censored at zero issues. These efforts demonstrate that multivariate models become a potential useful method for analysis when modeling the counts of specific types of crashes. However, as finer crash categorization became available, a significant amount of zeros appeared which are challenging the ability of the conventional multivariate regression model (i.e. MVP, multivariate negative binomial (MVNB), and MVPLN). As the crash data with excess zeros become an issue, there is a need of model development that can describe data characterized with a preponderance of zeros.

Zero-inflated models have been applied to cope with data characterized by a significant amount of zero observations or more zero observations than the one would expect in a traditional model (Poisson or negative binomial model). Zero-inflated models operate on the principle that the excess zeros is accounted for by a splitting regime that models a virtually safe state versus a crash-prone propensity of a roadway entity. The probability of an intersection being in zero or non-zero states can be determined by a binary logit or probit model (Lambert, 1992; Lee and Mannering, 2002; Kumara and Chin, 2003). Despite zero-inflated model has been used broadly under the situations where the observed data are characterized by large zero densities, Lord et al. (2005, 2007) have criticized the application of this model in highway safety. They argued that zero-inflated model cannot properly reflect the crash-data generating process, since the zero or safe state has a long-term mean equal to zero. Recently, the problems associated with a dual-state data generating process have been discussed by other researchers. As an alternative, Malyshkina and Mannering (2010) have proposed a zero-state Markov switching count-data model for circumventing such problem. Markov switching models offer considerable potential for providing important new insights into the analysis of crash data. However, these models are quite complex to estimate. Furthermore, there is no evidence that the Markov switching model can be extended to bivariate/multivariate formulation.

Though the zero-inflated model had limitations, we chose it as the baseline model based on the following considerations. First, literature indicates that zero-inflated models provide improved statistical fit compared to traditional Poisson and NB models (Lee and Mannering, 2002; Kumara and Chin, 2003; Shankar et al., 2003). Second, the crash data in our dataset were characterized by a preponderance of zeros, which is caused by one or more of the following conditions: (1) the subject of the study is the intersections, which represent small spatial scales; (2) analysis of intersections are characterized by a combination of low exposure, high heterogeneity, and sites categorized as high risk; (3) the number of zero observations increases as crash counts have many categories; and (4) it is possible that the sample data contain a relatively high percentage of non-reported crashes. To deal with the problems associated with using crash data characterized by a large number of zeros, we employed zero-inflated models as the best modeling approach. Nevertheless, univariate zero-inflated models cannot handle the correlation problem among specific types of crashes. So there is a need to develop a multivariate zero-inflated regression model to handle the situation which involves more than one type of crash and crash data were characterized by a significant amount of zeros.

In the current paper, a multivariate approach is introduced for jointly modeling data on crash counts by severity on the basis of MZIP distributions. Using this MZIP specification, as well as Bayesian estimation techniques, our study models correlated traffic crash counts simultaneously at different levels of severity by using crash data for signalized intersections in Tennessee. In addition, the paper investigates the performances of MVPLN, ZIP, and MZIP regression models in establishing the relationship between crashes, traffic factors, and geometric design features of roadway intersections.

3. Model structure and estimation

The following section presents the general forms of MZIP regression models and provides brief descriptions of its estimation procedures. Bivariate ZIP (BZIP) distributions are presented first, since we adopted the ideas of constructing the BZIP to construct the MZIP distribution.

3.1. BZIP distributions

There are at least three methods to construct a BZIP model (Li et al., 1999; Walhin, 2001; Wang et al., 2003). In this paper, the BZIP distribution is constructed as a mixture of a bivariate Poisson, two univariate Poisson, and a point mass at (0, 0), which has the property that the marginal distributions are univariate ZIP's. The probability mass function (pmf) of the BZIP is given by

$$F(y_1,y_2)=\begin{cases} p_0+p_1\exp(-\lambda_1)+p_2\exp(-\lambda_2)+p_{11}\exp(-\lambda) & y_1=y_2=0 \\ [p_i\lambda_i^{y_i}\exp(-\lambda_i)+p_{11}\lambda_{i0}^{y_i}\exp(-\lambda)]/y_i! & y_1 \text{ or } y_2\neq 0,\ \text{if } y_1\neq 0,\ \text{then } i=1,\ \text{otherwise } i=2 \\ p_{11}\sum_{j=0}^{\min(y_1,y_2)}\lambda_{10}^{y_1-j}\lambda_{20}^{y_2-j}\lambda_{00}^{j}\exp(-\lambda)/[(y_1-j)!(y_2-j)!j!] & y_1,y_2\neq 0 \end{cases} \quad (1)$$

where y_i is the number of crashes for crash type i and $\lambda=\lambda_{10}+\lambda_{20}+\lambda_{00}$. Note that Y_1 and Y_2 can be represented by $Y_1=U_1+U_0$ and $Y_2=U_2+U_0$, where U_1, U_2, and U_0 are independent univariate Poisson random variables with mean λ_{10}, λ_{20}, and λ_{00}, respectively.

The distribution can be interpreted as follows: An intersection is in a virtually safe state with probability p_0, and the intersection is in a crash-prone propensity state i with probability p_i, where the number of crashes has a Poisson distribution with mean λ_i.

3.2. MZIP distributions

This section extends the BZIP to an m-dimensional multivariate ZIP (MZIP). We follow the ideas of constructing the BZIP used in Section 3.1. First, a multivariate Poisson distribution is needed. There are several methods to construct a multivariate Poisson. One way to construct a m-dimensional Poisson is to use just one common term (Li et al., 1999),

$$Y_1=U_1+U_0, Y_2=U_2+U_0,\cdots,Y_m=U_m+U_0 \quad (2)$$

though the covariance structure of the preceding construction is limited compared to the more general definition of the multivariate Poissons, there are a manageable number of parameters to handle. For simplicity of the model in our presentation, we use this distribution as an m-dimensional Poisson to construct a MZIP model.

To obtain a general m-dimensional ZIP model, we use a mixture of a point mass at 0 (virtually safe state), m-distributions each with univariate Poisson for one crash type and $m-1$ zeros-for example, (Poisson (λ_1), 0,...,0),...,(0,...,0, Poisson (λ_m)), ($m(m-1)/2 = C_m^2$) distributions each with bivariate Poisson for two crash types and $m-2$ zeros,...,m distributions each with $(m-1)$-dimensional Poisson for $m-1$ crash types and one 0-and finally, an m-dimensional Poisson for all m crash type. Because MZIP distributions are used mainly for situations in which most crash counts are 0, we pool all of the mixtures of (m choose k) distributions with k-variate Poisson for k crash types and $(m-k)$ zeros, $m\geqslant k\geqslant 2$ together, and use the m-dimensional Poisson distribution to represent them. Let $\lambda=\lambda_{10}+\lambda_{20}+\cdots+\lambda_{m0}+\lambda_{00}$. The pmf of the MZIP is

$$F(y_1,y_2,\cdots,y_m)=\begin{cases} p_0+p_1\exp(-\lambda_1)+p_2\exp(-\lambda_2)+\cdots+p_m\exp(-\lambda_m)+p_{11}\exp(-\lambda) & \sum_{j=1}^{m}y_j=0 \\ [p_i\lambda_i^{y_i}\exp(-\lambda_i)+p_{11}\lambda_{i0}^{y_i}\exp(-\lambda)]/y_i! & y_i\neq 0 \text{ and } \sum_{j=1,j\neq i}^{m}y_j=0 \\ p_{11}\sum_{j=0}^{\min(y_1,y_2,\cdots,y_m)}\frac{\lambda_{10}^{y_1-j}\lambda_{20}^{y_2-j}\cdots\lambda_{m0}^{y_m-j}\lambda_{00}^{j}}{(y_1-j)!(y_2-j)!\cdots(y_m-j)!j!}\exp(-\lambda) & \text{at least two of the } y_i\text{'s are not 0} \end{cases} \quad (3)$$

where $p_0+p_1+p_2+\cdots+p_m+p_{11}=1$ and λ_{10}, $\lambda_{20},\ldots,\lambda_{m0}$, and λ_{00} are the means of U_1, $U_2,\ldots,U_m$, and U_0 in Eq. (2), respectively. The MIZP model has $2(m+1)$ parameters, which increase linearly with m. With the further assumption that $\lambda_1=\lambda_{10}+\lambda_{00}$, $\lambda_2=\lambda_{20}+\lambda_{00},\ldots,\lambda_m=\lambda_{m0}+\lambda_{00}$, it can be verified that any marginal distribution with m_1 variables, where $m_1<m$, is also m_1-dimensional MZIP. When (Y_1, $Y_2,\ldots,Y_m$) has an MZIP distribution (3), the marginal distribution of Y_i is

$$\begin{aligned} Y_i &\sim 0 && \text{with probability } 1-p_i-p_{11} \\ &\sim \text{Poisson}(\lambda_i) && \text{with probability } p_i+p_{11} \end{aligned} \quad (4)$$

Thus Y_i has a univariate ZIP distribution marginally and its mean and variance are $E(Y_i)=(p_i+p_{11})\lambda_i$ and $\text{var}(Y_i)=(p_i+p_{11})\lambda_i[1+(1-p_i-p_{11})\lambda_i]$, respectively.

3.3. MZIP regression model

If the joint probabilities in MZIP can be linked to the geometric variables and other attributes, one can build a regression model to characterize the geometric design features and use the model to find a way to control the geometric features to improve safety performance at intersections. Suppose that the observations $y_i=(y_{i1}, y_{i2},\ldots,y_{im})$ are conditionally independent across intersections. Define the following simple linear regression relationship to link the MZIP parameters to the geometric effect and $X_i=(1, x_{i1}, x_{i2},\ldots,x_{iK})'$, where x_{ik} is the geometric factor. The vectors of MZIP parameters p_{ij} and $\lambda_i=(\lambda_{i10}, \lambda_{i20},\ldots,\lambda_{im0}, \lambda_{i00})$ satisfy

$$\begin{aligned} &y_{ij}|\beta_j,\gamma_j\sim \text{MZIP}(\lambda_{ij0},p_{ij}) \\ &\lambda_{ij0}=\exp(\beta_j\mathbf{X}_i) \\ &p_{ij}=\exp(\gamma_j\mathbf{X}_i)/\left[1+\sum_{j=1}^{m}\exp(\gamma_j\mathbf{X}_i)\right] \text{ and } p_{11}=1-\sum_{j=1}^{m}p_{ij} \end{aligned} \quad (5)$$

where $\beta_j=(\beta_{j0}, \beta_{j1},\ldots,\beta_{jK})$, and $\gamma_j=(\gamma_{j0}, \gamma_{j1},\ldots,\gamma_{jK})$ are vectors of coefficients, $j=0, 1, 2,\ldots,m$.

4. Data description

Crash data obtained from the Tennessee Roadway Information System (TRIMS) are employed to evaluate the performance of the MZIP regression model described above in terms of its ability to establish the relationship between crashes, traffic factors, and geometric features. The analysis considers crashes occurring within the intersection or within 76 m (250 ft) of the center of the intersection along the major and minor roads. While this classification scheme may omit some intersection crashes and/or

include some non-intersection crashes, it is commonly used in the United States (Ye et al., 2009). The criterion is non-arbitrary, easily repeatable, and generalizable across jurisdictions.

The study time increment is one year, which means that the same intersection, even if nothing had changed, is considered as five independent sections—one for each year from 2005 to 2009. This allows the model to consider year-to-year changes in intersection geometric design, traffic conditions, and other relevant attributes.

The MZIP model is applied to crash count data of five different severity levels—Sev1, killed/fatal (K); Sev2, incapacitating injury (A); Sev3, non-incapacitating injury (B); Sev4, possible-injury (C); and Sev5, property damage only (PDO or O). The final data set includes data for 1000 intersections, for which a total of 22,079 crashes were reported over the 5-years period. There were 205 Sev1 crashes, 586 Sev2 crashes, 5398 Sev3 crashes, 14,193 Sev4 crashes, and 1697 Sev5 crashes at those 1000 intersections for 5 years, with an approximately 0.93%, 2.65%, 24.45%, 64.28%, and 7.69% split of fatal, incapacitating injury, non-incapacitating injury, possible injury, and PDO crashes, respectively. Individual intersections experienced from 0 to 39 crashes per year. As expected, a significant amount of zeros is observed. The proportions of intersections that experience fatal-crash free, incapacitating-injury-crash free, non-incapacitating-injury-crash free, possible-injury-crash free, and PDO-crash free are 96.34%, 90.23%, 64.81%, 56.24%, and 81.43%, respectively.

A series of explanatory variables describing intersection geometry characteristics and traffic factors for both major and minor roads were obtained from roadway characteristics inventories and GIS roadway networks and merged with the crash frequency database. Traffic volumes varied widely from intersection to intersection: from about 3386 to 138,525 vehicles per day for major approaches (summing over both directions) and from 1580 to about 99,975 vehicles per day for minor approaches. A descriptive analysis of the database is presented in Table 1.

5. Model estimation results

The WinBUGS software (Spiegelhalter et al., 2005)—an interactive Windows version of the BUGS program for Bayesian analysis of complex statistical models using MCMC techniques—was employed in this study for coefficient estimation. The techniques generate sequences (chains) of random points, whose distributions converge to the target posterior distributions. The estimations from burn-in iterations to convergence have been excluded. The remaining iterations are used for parameter estimation, performance evaluation, and inference. To ensure that the chain has converged to the posterior distribution by the end of the burn-in period, trace plots and the autocorrelation function plots of posterior sample values were obtained. All predicted values of concerned coefficients fall in the extent that does not produce strong periodicities and tendencies, in other words, the model converges.

5.1. Goodness of fit and model evaluation

For comparison, we applied a MVPLN model and five individual univariate ZIP models to fit the data. A deviance information criterion (DIC)—as proposed by Speigelhalter et al. (2003)—can assess the statistic fit of the model, was employed in this study to evaluate the goodness of fit. The last column of Table 2 gives the DIC statistic of each model. Based on the DIC statistics, the MZIP model fits the data significantly better than the MVPLN model. Fitting five univariate ZIP models result in a sum of DIC statistic of 4148.4, which is significantly greater than the DIC statistic 2086.3 from the MZIP model. This provides empirical support for the MZIP model over univariate ZIP model.

In addition, out-of-sample predictions from MZIP, univariate ZIP, and MVPLN model are compared for the different crash groups categorized by severity. Our data set contained 1000 roadway intersection samples, and examined a one year period. Of the sampled intersections, 965 had no Sev1 crashes, 890 had no Sev2 crashes, 633 had no Sev3 crashes, 580 had no Sev4 crashes, and 819 had no Sev5 crashes. In the 35 Sev1 crash intersection samples, 19 (54.84%) have only Sev1 crashes. Sev2 crashes occurred exclusively in only 31 (28.57%) of the 110 sample intersections. Just 101 (27.61%) of 367 intersection samples have Sev3 only crashes. Sev4 crashes occurred exclusively in only 135 (32.17%) of the 420 sample intersections. Just 43 (23.60%) of 181 intersection samples have Sev5 only crashes. Those statistics infer that different intersection features contribute differences in crash frequencies across severity level. Given a set of highway geometric design variables, determining which variables are relatively more critical to specific type of crash counts is important.

Table 2 suggests that the MZIP model predicts better than the MVPLN and univariate ZIP model. We hypothesize that the MZIP model better addresses the issue of unobserved heterogeneity

Table 1
Descriptive statistics of variables for Tennessee intersection data.

Variable name	Variable description	Mean	Std. err.	Min.	Max.
Dependent variables					
Sev1	Number of fatal crashes	0.041	0.221	0	3
Sev2	Number of incapacitating crashes	0.117	0.486	0	4
Sev3	Number of non-incapacitating crashes	1.080	2.171	0	21
Sev4	Number of possible injury crashes	2.841	3.704	0	34
Sev5	Number of PDO crashes	0.340	0.928	0	13
Independent variables					
AADTMAJ	Logarithm of AADT on major road	4.601	0.266	3.530	5.142
AADTMIN	Logarithm of AADT on minor road	4.279	0.388	3.199	5.000
PTRUCKMAJ	Percentage of trucks in traffic stream on major road (%)	10.009	2.636	1	17
PTRUCKMIN	Percentage of trucks in traffic stream on minor road (%)	7.413	4.314	0	15
LWIDTHMAJ	Through lane width on major road	11.010	1.000	10	12
LWIDTHMIN	Through lane width on minor road	11.010	0.702	10	12
SWIDTHMAJ	Shoulder width on major road (ft)	3.323	2.983	0	10
SWIDTHMIN	Shoulder width on minor road (ft)	3.810	3.345	0	11
LLANEMAJ	Number of left-turn lane on major road	0.814	0.541	0	2
LLANEMAIN	Number of left-turn lane on minor road	0.497	0.500	0	2
SLIMITMAJ	Posted speed limit on major road (mph)	45.371	7.962	25	65
SLIMITMIN	Posted speed limit on minor road (mph)	42.497	10.366	25	60
ANGLE	Angle of intersection (degree)	75.068	8.715	60	90
No. of crashes		22,079			

Table 2
Comparisons of predictions from MZIP, ZIP, and MVPLN models for 1000 intersections in 2010.

	P^*	Fatal crashes	Incapacitating crashes	Non-incapacitating crashes	Possible injury crashes	PDO crashes	Goodness of fit (DIC)
Observed	0.3649	38	98	998	2465	328	
MZIP model	0.3343 (−8.40%)	41 (7.89%)	105 (6.92%)	1047 (4.95%)	2566 (4.10%)	345 (5.18%)	2086.3
Individual ZIP model	0.2989 (−18.09%)	45 (18.42%)	112 (14.05%)	1118 (12.03%)	2707 (9.82%)	382 (16.44%)	4148.4**
MVPLN model	0.3002 (−17.73%)	43 (13.16%)	110 (12.22%)	1104 (10.63%)	2666 (8.16%)	367 (11.88%)	2247.6

Note: The values in parentheses represent the percentage difference (%) between the observed and prediction values.
* $P^* = P(Y_1 = 0, Y_2 = 0, Y_3 = 0, Y_4 = 0, Y_5 = 0)$.
** This is a combination DIC. The individual ZIP DIC statistics were 1167.9, 836.7, 707.8, 701.7, and 734.3 for fatal, incapacitating, non-incapacitating, possible injury, and PDO crash models, respectively.

Table 3
Posterior summary of Bayesian MZIP model fitting—β.

Variable name	β_0	Fatal crashes	Incapacitating crashes	Non-incapacitating crashes	Possible injury crashes	PDO crashes
Constant	−9.003* 0.064** (−9.012, −8.978)***	−11.206 0.046 (−11.231, −11.197)	−6.573 −0.015 (−6.602, −6.556)	−2.727 0.049 (−2.913, −2.650)	−1.012 0.023 (−1.128, −0.709)	−5.565 0.014 (−6.947, −4.935)
AADTMAJ	1.024 0.019 (1.018, 1.029)	0.634 0.034 (0.597, 0.649)	0.218 0.036 (0.208, 0.236)	0.485 0.008 (0.458, 0.496)	0.305 0.023 (0.301, 0.338)	0.547 0.037 (0.535, 0.551)
AADTMIN	3.014 0.016 (2.997, 3.016)	0.898 0.023 (0.896, 0.913)	0.820 0.032 (0.784, 0.836)	0.534 0.024 (0.521, 0.542)	0.320 0.030 (0.282, 0.359)	0.692 0.022 (0.675, 0.729)
PTRUCKMAJ	0.207 0.031 (0.191, 0.220)	0.108 0.034 (0.092, 0.112)	0.172 0.022 (0.169, 0.198)	–	–	–
PTRUCKMIN	0.010 0.414 (0.404, 0.442)	0.178 0.004 (0.150, 0.180)	0.071 0.033 (0.051, 0.084)	–	–	–
LLANEMAJ	1.109 0.031 (1.084, 1.121)	–	–	0.033 0.030 (0.004, 0.054)	−0.061 0.006 (−0.078, −0.043)	−0.281 0.027 (−0.303, −0.255)
LLANEMAIN	−0.098 0.027 (−0.136, −0.063)	0.112 0.012 (0.074, 0.137)	0.107 0.024 (0.105, 0.117)	0.101 0.025 (0.083, 0.117)	−0.088 0.016 (−0.115, −0.072)	−0.055 0.022 (−0.058, −0.030)
ANGLE	−0.159 0.012 (−0.164, −0.123)	−0.031 0.034 (−0.063, −0.025)	−0.035 0.022 (−0.036, −0.004)	−0.024 0.002 (−0.048, −0.003)	−0.009 0.032 (−0.049, −0.002)	−0.011 0.007 (−0.024, −0.006)

* Posterior means.
** Posterior standard deviations.
*** 95% Credible intervals.

and allows for correlations among crash counts across severity. Results in Table 2 show that the crash-free probability from the MZIP model are very close to the observed proportion of zeros (36.49%). The predicted percentages of fatal-crash free, incapacitating-injury-crash free, non-incapacitating injury-crash free, possible-injury-crash free, and PDO-crash free are 97.19%, 91.12%, 63.31%, 54.83%, and 83.65% in MZIP model, respectively, which are very close to the observed crash-free proportion in the specific type of crash samples. The corresponding P (30.02%) for the MVPLN is underestimated. Treating crash type independently and computing the crash-free probability as $P(Y_1 = 0)P(Y_2 = 0)P(Y_3 = 0)$-$P(Y_4 = 0)P(Y_5 = 0)$, obtained from individual ZIP models, results in a 29.89% estimate of virtually safe state probability, which is also an underestimate compared to the MZIP and the true proportion of zero crashes.

5.2. Estimation results

Tables 3 and 4 summarize the estimations (posterior means, standard deviations, and 95% credible intervals) of the regression coefficients β and γ based on the MIZP model. The effects of significant factors on crash frequencies are interpreted by crash counts across severity.

Traffic volume and intersection angle are identified as significant factors for all crash types. Traffic volume contributes positively to the frequencies of all crash types while intersection angle (from less than 90° to 90°) contributes negatively. Those are expected results since increasing traffic volumes mean higher likelihood of crash occurrences and crossing a skewed-angle intersection results in an increased exposure time to the cross street traffic.

The percentage of truck is identified as a significant factor in fatal and incapacitating crash models. The fatal and incapacitating crash frequency increases as the percentage of truck increases. We believe that this occurs because the chance of a collision with a truck increases as the proportion of trucks in a traffic stream of a particular volume increases.

The number of left turn lanes on the major roads is identified as a significant factor for non-incapacitating, possible injury, and PDO crash frequencies. The results indicate that the number of left turn lanes on the major roads positively impacts non-incapacitating crash frequencies. This is consistent with the observation that sideswipe collisions are more likely to occur when transitioning into

Table 4
Posterior summary of Bayesian MZIP model fitting—γ.

Variable name	γ_0	Fatal crashes	Incapacitating crashes	Non-incapacitating crashes	Possible injury crashes	PDO crashes
Constant	−7.074[*] 0.059[**] (−7.087, −7.054)[***]	3.108 0.024 (3.084, 3.126)	4.04 0.02 (4.018, 4.065)	3.308 0.038 (3.083, 3.429)	−1.317 0.014 (−1.329, −1.217)	2.136 0.028 (1.006, 3.054)
AADTMAJ	−0.743 0.031 (−0.779, −0.739)	−0.013 0.009 (−0.048, −0.010)	−0.352 0.032 (−0.380, −0.334)	−0.742 0.003 (−0.780, −0.715)	−0.592 0.008 (−0.613, −0.573)	−0.751 0.018 (−0.773, −0.724)
AADTMIN	−0.864 0.027 (−0.877, −0.856)	−0.011 0.028 (−0.018, −0.005)	−0.134 0.01 (−0.156, −0.113)	−1.514 0.013 (−1.532, −1.505)	−1.878 0.005 (−1.918, −1.843)	−0.533 0.003 (−0.559, −0.529)
PTRUCKMAJ	−0.178 0.015 (−0.184, −0.166)	−0.105 0.03 (−0.118, −0.080)	−0.465 0.007 (−0.495, −0.456)	–	–	–
PTRUCKMIN	−0.169 0.008 (−0.172, −0.162)	−0.779 0.013 (−0.806, −0.761)	−0.496 0.011 (−0.526, −0.484)	–	–	–
LLANEMAJ	0.555 0.031 (0.553, 0.591)	–	–	−0.684 0.025 (−0.694, −0.656)	0.134 0.008 (0.099, 0.166)	0.023 0.021 (0.008, 0.035)
LLANEMAIN	0.095 0.034 (0.074, 0.100)	−0.386 0.011 (−0.425, −0.369)	−0.603 0.002 (−0.616, −0.591)	−0.136 0.024 (−0.155, −0.123)	0.122 0.009 (0.100, 0.152)	0.052 0.027 (0.046, 0.053)
ANGLE	0.036 0.027 (0.025, 0.045)	0.027 0.026 (0.011, 0.064)	0.103 0.028 (0.088, 0.131)	0.131 0.003 (0.107, 0.133)	0.197 0.009 (0.175, 0.202)	0.07 0.013 (0.037, 0.087)

[*] Posterior means.
[**] Posterior standard deviations.
[***] 95% Credible intervals.

Table 5
Expected percentage changes in crash frequencies corresponding to changes in variables.

Variable	Average	Change in variable	Percentage change in crash frequencies (per year)				
			Fatal crashes (%)	Incapacitating crashes (%)	Non-incapacitating crashes (%)	Possible injury crashes (%)	PDO crashes (%)
AADTMAJ	4.601	0.25	17.16	5.61	12.89	7.92	14.66
AADTMIN	4.279	0.25	25.18	22.76	14.28	8.33	18.90
PTRUCKMAJ	10	1	11.40	18.76	–	–	–
PTRUCKMIN	7	1	19.54	7.37	–	–	–
LLANEMAJ	1	1	–	–	3.37	−5.88	−24.50
LLANEMAIN	0	1	11.81	11.30	10.65	−8.41	−5.34
ANGLE	75	10	−26.85	−29.31	−21.46	−8.70	−10.24

left turn lanes at an intersection. The number of left turn lanes on the minor roads leads to more fatal and incapacitating crashes. Minor roads often do not have protected left turn signal indications, even when the intersection is signalized. As the presence of left turn lanes is likely an indicator of a higher level of left turning movements (from the minor roads), this can contribute to a higher number of angle crashes. In addition, at the beginning of the red phase, there will be a higher likelihood of conflict with permissive movements on the major roads and RLR (red-light running) on the minor roads.

The results also show that the presence of left turn lanes on both major and minor roads reduces the possible injury and PDO crash frequencies, but their effect varies across the severity level. One possible explanation for this finding is that the presence of left turn lanes may result in crashes of lower severity, and because the data used consists only of police-reported crashes, this finding could reflect, among other factors, under-reporting of minor crashes.

Based on an average roadway intersection's attributes and the MZIP model's average parameter estimations, Table 5 provides estimations of percentage changes in crash frequencies as a function of various design features. For example, adding one left lane on a minor road (from 0 to 1) is predicted to result in 11.81%, 11.30%, and 10.65% greater fatal, incapacitating, and non-incapacitating crashes, respectively, and to decrease the number of possible injury and PDO crashes by 8.41% and 5.34%.

Table 5 reveals that the number of left turn lanes on major roads appears to significantly affect non-incapacitating, possible injury, and PDO crash frequencies, but have a slight effect on fatal and incapacitating crash frequencies. The results indicate that reducing the number of left turn lanes on the major roads will lead to more possible injury or PDO crashes. Thus, if an intersection experiences high possible injury or PDO crash frequencies and other types of crashes are not an issue, increasing the left-turn lane number on the major road may be an effective countermeasure. In addition, increasing the number of left lanes on the major roads increases non-incapacitating crash frequencies, but will not increase fatal and incapacitating crashes significantly. Another finding suggests that increasing the number of left turn lanes on the minor roads increases fatal, incapacitating, and non-incapacitating crash frequencies, but has little effect on possible injury and PDO crashes. Improving the angle of intersection can be considered as the effective countermeasures for intersections with a high rate of all crash types.

6. Conclusions

In modeling roadway crashes, zero-inflated count models have been shown potential useful in obtaining a better model-fitting when zero crash counts are over-presented. However, the general specification of zero-inflated regression models cannot handle the data on crash frequency by severity since they are multivariate in nature. A multivariate approach for modeling crash data by severity is presented in this paper based on the MZIP distributions. The MCMC algorithm was employed as a computational engine for the parameter estimations. The method was applied to the multivariate crash counts from 1000 intersections in Tennessee obtained for 5 years. The findings suggest that traffic volume contributes positively to the frequencies of all crash types. Examining the effects of car-truck mix suggests that, as the percentage of truck increases, fatal and incapacitating crash frequencies increase significantly. In addition, increases in intersection angle degree significantly decreases fatal, incapacitating, non-incapacitating, possible injury, and PDO crash frequencies. The number of left turn lanes on major roads contributes positively to non-incapacitating crash frequencies, while having a negative impact on the possible injury and PDO crash frequencies. The number of left turn lanes on the minor roads positively impacts fatal, incapacitating, and non-incapacitating crash frequencies, while having a negative impact on the possible injury and PDO crash frequencies.

As expected, there were excess zeros in the correlated crash count data that cannot be handled by either the univariate ZIP models nor the previously suggested MVPLN models. The empirical results reveal that the MZIP model outperforms its univariate counterpart and is practically superior to the MVPLN model. The MZIP model, compared to the univariate ZIP and MVPLN model, provides the best model fit and has the smallest prediction bias. Apart from the improvement in goodness of fit, the results of MZIP show promise toward the goal of obtaining more accurate estimates by accounting for excess zeros in correlated count data. Two suggestions are drawn from this research, as summarized below.

(1) It is important to use the MZIP model to fit the crash data when there is a high probability that the intersections are in a virtually safe state and the variables are correlated.
(2) In this paper, the Bayesian methodologies have been used to estimate the parameters by using sampling-based methods. From this study, it is evident that the proposed methods are quite effective in drawing inferences based on small samples.

It can be seen that the proposed method can be implemented easily and also it can be extended to other multivariate zero-inflated distributions involving other probability distributions (e.g., zero-inflated negative binomial distribution). However, MZIP models are very complex to estimate. In addition, though MZIP models can account for correlation between crash types, there could be correlation within the sites that needs to be estimated. Cluster sampling methods or random-parameters models would offer considerable potential for providing important new insights into the analysis of crash data with regard to this issue.

Acknowledgements

The authors would like to thank both the editor and reviewers for their patience and numerous thoughtful comments which improved the quality of the paper dramatically. In addition, special thanks to TDOT for providing the TRIMS data. This research is supported by the Southeastern Transportation Center – a Regional UTC funded by the USDOT Research and Innovative Technology Administration. Additional funding was provided by the Natural Science Foundation of China (No. 51208052).

References

Anastasopoulos, P., Shankar, V., Haddock, J., Mannering, F., 2012. A multivariate tobit analysis of highway accident-injury-severity rates. Accid. Anal. Prev. 45, 110–119.
Bijleveld, F.D., 2005. The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accid. Anal. Prev. 37 (4), 591–600.
Chang, L., Chen, W., 2005. Data mining of tree-based models to analyze freeway accident frequency. J. Safety Res. 36 (4), 365–375.
Dong, C., Clarke, D.B., Richards, S.H., Huang, B., 2014. Differences in passenger car and large truck involved crash frequencies at urban signalized intersections: an exploratory analysis. Accid. Anal. Prev. 62, 87–94.
EI-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. Accid. Anal. Prev. 41 (4), 820–828.
Kumara, S.S.P., Chin, H.C., 2003. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. Traffic Injury Prevent. 3 (4), 53–57.
Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34 (1), 1–14.
Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accid. Anal. Prev. 34 (2), 149–161.
Li, C., Lu, J., Park, J., Kim, K., Brinkley, P., Peterson, J., 1999. Multivariate zero-inflated Poisson models and their applications. Technometrics 41 (1), 29–38.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A 44 (5), 291–305.
Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37 (1), 35–46.
Lord, D., Washington, S., Ivan, J.N., 2007. Further notes on the application of zero-inflated models in highway safety. Accid. Anal. Prev. 39 (1), 53–57.
Ma, J., Kockelman, K.M., 2006. Bayesian multivariate Poisson regression for models of injury count, by severity. Transport. Res. Record: J. Transport. Res. Board No. 1950, 24–34.
Ma, J., Kockelman, K., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity using Bayesian methods. Accid. Anal. Prev. 40 (3), 964–975.
Malyshkina, N.V., Mannering, F.L., 2010. Zero-state Markov switching count-data models: an empirical assessment. Accid. Anal. Prev. 42 (1), 122–130.
Miaou, S., Song, J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion and spatial dependence. Accid. Anal. Prev. 37 (4), 699–720.
Park, E., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transport. Res. Record: J. Transport. Res. Board No. 2019, 1–6.
Shankar, V.N., Ulfarsson, G.F., Pendyala, R.M., Nebergall, M.B., 2003. Modeling crashes involving pedestrians and motorized traffic. Safety Sci. 41 (7), 627–640.
Song, J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. J. Multivariate Anal. 97 (1), 246–273.
Speigelhalter, D., Best, N., Carlin, B., van der Linde, A., 2003. Bayesian measures of model complexity and fit. J. Roy. Stat. Soc.: Ser. B (Stat. Methodol.) 64 (4), 583–616.
Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., 2005. WinBUGS User Manual. MRC Biostatistics Unit, Cambridge. Available from <http://www.mrc-cam.ac.uk/bugs>.
Walhin, J., 2001. Bivariate ZIP models. Biometrical J. 43 (2), 147–160.
Wang, K., Lee, A., Yau, K., Carrivick, P., 2003. A bivariate zero-inflated Poisson regression model to analyze occupational injuries. Accid. Anal. Prev. 35 (4), 625–629.
Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Safety Sci. 47 (3), 443–452.
Zhang, Y., Xie, Y., Li, L., 2012. Crash frequency analysis of different types of urban roadway segments using generalized additive model. J. Safety Res. 43 (2), 107–114.