Analytic Methods in Accident Research 8 (2015) 45–60

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

An efficient parallel sampling technique for Multivariate Poisson-Lognormal model: Analysis with two crash count datasets

Xianyuan Zhan, H.M.Abdul Aziz, Satish V. Ukkusuri *

Lyles School of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN 47907, USA

ARTICLE INFO

Article history:
Received 25 November 2014
Received in revised form
27 October 2015
Accepted 27 October 2015

Keywords:
Accident analysis
Pedestrian crashes
Severity models

ABSTRACT

This study investigates the Multivariate Poisson-lognormal (MVPLN) model that jointly models crash frequency and severity accounting for correlations. The ordinary univariate count models analyze crashes of different severity level separately ignoring the correlations among severity levels. The MVPLN model is capable to incorporate the general correlation structure and also takes account of the overdispersion in the data that leads to a superior data fitting. However, the traditional estimation approach for MVPLN model is computationally expensive, which often limits the use of MVPLN model in practice. In this work, a parallel sampling scheme is introduced to improve the original Markov Chain Monte Carlo (MCMC) estimation approach of the MVPLN model, which significantly reduces the model estimation time. Two MVPLN models are developed using the pedestrian–vehicle crash data collected in New York City from 2002 to 2006, and the highway-injury data from Washington State (5-year data from 1990 to 1994) The Deviance Information Criteria (DIC) is used to evaluate the model fitting. The estimation results show that the MVPLN models provide a superior fit over univariate Poisson-lognormal (PLN), univariate Poisson, and Negative Binomial models. Further, the correlations among the latent effects of different severity levels are found significant in both datasets,that justifies the importance of jointly modeling crash frequency and severity accounting for correlations.

1. Introduction

Analyzing road crash-related injuries by severity level is critical for design and implementation of traffic safety countermeasures. The estimated cost of crashes can be significantly different at different severity levels, which poses great impact in the design of safety intervention techniques. For instance, a road segment with higher frequency of fatal crashes is more hazardous than a road segment with fewer fatal crashes but with more injury crash occurrences (Wang et al., 2011). From modeling perspective, the crash count data at different severity levels are potentially correlated due to both observed and unobserved factors (Mannering and Bhat, 2014). Multivariate count models offer estimation methods that allow for the correlation among severity levels. Studies analyzing the frequency of crashes and the resulting severity levels are affluent in the literature. With few exceptions, most studies apply univariate models without accounting for the possible correlations among different severity levels (e.g., the correlation between number of fatal and severe injury crashes). Significant

* Corresponding author. Tel.: +1 765 494 2296; fax: +1 765 496 7996.
E-mail addresses: zhanxianyuan@purdue.edu (X. Zhan), aziz.husain.nexus@gmail.com (H.M. Abdul Aziz), sukkusur@purdue.edu (S.V. Ukkusuri).

http://dx.doi.org/10.1016/j.amar.2015.10.002
2213-6657/

correlation among different crash related outcomes has already been reported in the literature (Bijleveld, 2005), which can be either caused by shared information in latent effects (Chib and Winkelmann, 2001), spatially or temporal dependencies (Aguero-Valverde and Jovanis, 2010; Barua et al., 2014), or omitted variables that impact all levels of crash frequencies (El-Basyouny and Sayed, 2009). Further, univariate models that focus on specific crash severity level also suffer from similar shortcomings. Both approaches treat the correlated crash counts as independent and accordingly lead to a less accurate or sometimes even misspecified model estimation result (Park and Lord, 2007). As a result, insights from these models without the aforementioned considerations can readily lead to ineffective countermeasures and safety policies. Multivariate count models that simultaneously analyze different severity levels using the same set of explanatory variables can overcome the limitations discussed above (Pei et al., 2011; Mannering and Bhat, 2014). This also solves the generalization problem when inferring from risk factors associated with the crashes (e.g. the effects of variables are not generalized for different levels of severity, that same variable might have different impacts on different severity levels). Unlike the field of long matured univariate count models, the field of multivariate count data models is relatively new. Most of the multivariate count models in literature were constructed as multivariate generalization and variation of Poisson class of models, such as Multivariate Poisson (MVP) model (Kocherlakota and Kocherlakota, 1992), Multivariate Negative Binomial (MVNB) model (Winkelmann, 2000; Caliendo et al., 2013), Poisson-Gamma Mixture model (Hausman et al., 1984) and Multivariate Poisson Log-Normal (MVPLN) model (Park and Lord, 2007; Ma et al., 2008; El-Basyouny and Sayed, 2009; Aguero-Valverde and Jovanis, 2009; Aguero-Valverde, 2013). Another major branch of multivariate count models focus on modeling the spatial and temporal dependencies in multi-severity crash data (Song et al., 2006; Aguero-Valverde and Jovanis, 2010; Castro et al., 2012; Narayanamoorthy et al., 2013; Wang and Kockelman, 2013; Barua et al., 2014). Although aforementioned models are advantageous to account for correlation of crash counts among different severity levels, they are either too restrictive in modeling assumptions, relatively time-consuming to estimate, or have feasibility issue in the case of high dimensionality when encountered for spatial and temporal dependencies (Mannering and Bhat, 2014).

Multivariate Poisson (MVP), Multivariate Negative Binomial (MVNB) and Multivariate Poisson Log-Normal (MVPLN) models are among the most frequently used multivariate count models that account for Poisson variation and heterogeneity in the context of safety analysis. The MVP model has Poisson distribution as its marginal distribution, however, assumes the covariance for different severity levels to be identical and nonnegative. This is a highly restrictive assumption since different severity levels can have different covariance and the possibility of negative correlations cannot be discarded entirely (Park and Lord, 2007; Ma et al., 2008; El-Basyouny and Sayed, 2009). To relax the "equal covariance" assumption and account for overdispersion inherited from Poisson distribution, the Multivariate Negative Binomial (MVNB) and Poisson-Gamma Mixture models were developed. However, both models only allow for positive correlation, which still lacks a general correlation structure. Due to the limitation of previous models, the MVPLN approach is generally preferred over the previous approaches. Two key reasons are: (a) MVPLN can account for over-dispersion in the count data; (b) a general correlation structure can be used (Chib and Winkelmann, 2001; Park and Lord, 2007; El-Basyouny and Sayed, 2009). Although MVPLN models provide a sound framework for crash analysis, only a few studies applying the MVPLN model can be found in the literature (Park and Lord, 2007; Ma et al., 2008; El-Basyouny and Sayed, 2009; Aguero-Valverde and Jovanis, 2009; Aguero-Valverde, 2013). One major reason of the limited use of MVPLN model is associated with the difficulty in the model estimation. Previous studies developed multiple codes and tools to estimate MVPLN model, however, these tools were either failed to provide comprehensive goodness-of-fit measure or took considerable amount of time in model estimation, which limits the implementation of MVPLN model by practitioners.

This study implements the MVPLN framework on two crash datasets to demonstrate the superiority of MVPLN model on estimation accuracy and the ability to capture correlations among crash severity levels. The primary dataset is a comprehensive bivariate (fatal and severe injury) pedestrian–vehicle crash data from 2183 census tracts of New York City (NYC). A second and smaller highway injury dataset from 275 roadway segments in Washington State with three severity levels (no-injury, possible injury and evident injury) is also included to demonstrate the applicability of the MVPLN model and solution approach on multivariate case. To address the cumbersome and time-consuming computation issue of the original Monte Carlo Markov Chain (MCMC) estimation approach for MVPLN model, this study proposes a parallel sampling scheme for the original MCMC estimation approach and develops an efficient MATLAB codes to estimate the MVPLN models. The improved estimation approach significantly reduces the computation time for MVPLN models, which makes MVPLN model a more appealing analytical tool for practitioners to solve real world problems. A comparison study on MVPLN, univariate Poisson Log-Normal (PLN), univariate Poisson and Negative Binomial is also conducted to compare the model fitting results for different models. Finally, we also quantify and examine the extent of correlation among different crash severity levels from the two estimated models. In summary, this work contributed to the literature in the following aspects:

- Proposed a parallel sampling scheme to allow for efficient estimation of MVPLN model.
- Apply MVPLN framework to jointly model different crash severity levels using two different multivariate crash count dataset and providing insights about contributing factors of crashes.
- Revealed universal high level of correlations among different severity levels of crash counts in two types of crashes: urban pedestrian–vehicle crashes and highway vehicle crashes.
- Built a computationally efficient tool for practitioner to better apply MVPLN models.

2. Previous research

The conventional univariate count data models such as Poisson, Negative Binomial, Zero-Inflated Poisson and Zero-Inflated Negative Binomial models are widely used in crash analysis when modeling the total number of crashes (which include all severity outcomes). However, univariate count data models may not be appropriate when modeling individual severity levels because they do not account for possible correlation among the number of crashes occurring in each severity level. In this case, multivariate count data models are more appropriate for crash analysis that simultaneously modeling different severity levels and accounting for generalized correlation structure among the severity levels.

A number of multivariate count models that account for Poisson variation and heterogeneity in crash data have been developed in literature. Kocherlakota and Kocherlakota (1992) proposed the Multivariate Poisson (MVP) model, which uses Poisson distribution as its marginal distribution. Due to the ease of the estimation procedure, the MVP model became the only multivariate count data model that put to practical use in econometrics. However, MVP models cannot account for negative correlation and overdispersion in severity levels of each observation. As an extension of MVP, the Multivariate Negative Binomial (MVNB) model (Winkelmann, 2000; Caliendo et al., 2013) and Multivariate Poisson-Gamma Mixture model (Hausman et al., 1984) can account for overdispersion, but still only allow for non-negative correlations. The reason for inability to capture the negative correlation in the three abovementioned models is associated with their so-called one-factor structure, that the correlation is generated through an individual specific random factor which does not vary over outcomes (Winkelmann, 2008). To overcome the drawbacks of MVP, MVNB and Multivariate Poisson-Gamma Mixture models, Chib and Winkelmann (2001) developed the MVPLN model. The MVPLN models allow for a more general correlation structure (both positive and negative), and are also capable of capturing the overdispersion in the data.

Several studies have used MVPLN models to perform multivariate analysis for crash count data. Park and Lord (2007) developed MATLAB codes for MVPLN model to analyze the crash data with five different severity levels collected from 451 three-leg non-signalized intersections, and more accurate estimates were observed. Ma et al. (2008) coded an R program to model the crash counts by severity levels using data collected from Washington State through the HSIS on 7773 rural two-lane roadways in Puget Sound region. Their results showed that the MVPLN model provides better predictions than those from univariate Poisson and negative binomial models. Aguero-Valverde and Jovanis (2009) used full Bayes estimation approach on OpenBUGS to estimate MVPLN model. Compared with univariate Poisson lognormal (PLN) estimates, they showed that the MVPLN model fitted better than the univariate model and the correlations among crash severities are found high between contiguous severity levels. El-Basyouny and Sayed (2009) used the WinBUGS platform to model 99 signalized intersections in the city of Edmonton. Highly significant correlation between no-injury crashes and injuries plus fatalities (I+F) crashes was observed. Also, the results showed that MVPLN can lead to a higher precision in highly correlated no-injury and I+F crashes. Park et al. (2010) generalized the MVPLN model to encompass a change-point model that can analyze before-after data with comparison groups. The model was applied to Korean expressway crash data to perform safety evaluation of decreasing the speed limit, which showed the flexibility of the MVPLN modeling framework. All of these previous studies confirmed that MVPLN model provides superior fit to the data compared with univariate models.

Other multivariate modeling frameworks that generalize Poisson count models also exist in multivariate crash count data analysis. Ye et al. (2009) proposed a simultaneous equations model of crash frequencies by collision type using crash data for rural intersections in Geogia. However, it was reported that the results of the proposed model do not differ substantially than the MVP model, thus from a statistical standpoint, the gain in efficiency and goodness-of-fit was modest. Nikoloulopoulos and Karlis (2010) proposed a regression copula-based model for bivariate count data. Although this model offers some flexibility by being able to incorporate different marginal distributions, a limitation for this approach is that it restricts the application only to bivariate count data. Pei et al. (2011) proposed two specific joint-probability approaches under Bayesian framework to simultaneously modeling crash frequencies of different severity levels. However, this study only assessed a binary-severity case study at signalized intersections in Hong Kong and the performance of this model on multivariate crash count data still needs further examination. Chiou and Fu (2013) proposed a multinomial-generalized Poisson model and analyzed the accident for Taiwan's No. 1 Freeway. However, this model uses shared error component to handle the common error term and covariance structure, which is less flexible compared with the MVPLN model.

Besides the aforementioned models, another major class of multivariate models were developed to account for the spatial and temporal dependencies in multi-severity crash data (Song et al., 2006; Aguero-Valverde and Jovanis, 2010; Castro et al., 2012; Narayanamoorthy et al., 2013; Wang and Kockelman, 2013; Barua et al., 2014). Since the scope of this paper is not to investigate the spatial and temporal correlations in crash data, thus the detailed discussion of these models is omitted here.

Due to the nice features of the MVPLN model showed in literature, this study focuses on implementing MVPLN model to simultaneously model crash data on two dataset: a pedestrian–vehicle fatal and severe injury crash data in New York City, and a supplementary smaller scale highway injury data with three severity levels (no-injury, possible injury and evident injury) from Washington State. This study improves the original estimation procedure of MVPLN model by introducing a new parallelized MCMC scheme to speed up the model estimation. For comparison purpose, univariate Poisson-Lognormal (PLN), Poisson and Negative Binomial models were also estimated using the same set of explanatory variables and reported in the paper.

3. Methodology

3.1. The MVPLN model

The MVPLN assumes multivariate normal distributed latent effects in a Poisson distribution. Let y_{is} be the pedestrian–vehicle crash counts of census tract i ($i = 1, 2, \dots, n$) and severity level s ($s = 1, 2, \dots, S$). Let $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{is})^T$ be the severity-level specific latent effects for census tract i, and denote $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ the severity-level-specific latent effects across census tracts. Let x_{is} be the explanatory variables, and β_s be their coefficients. Assume the crash counts y_{is} conditioned on ε_i, the explanatory variables x_{is}, and corresponding coefficients β_s are independent Poisson distributed:

$$y_{is}|\varepsilon_i, \beta_s, x_i \sim Poisson(\lambda_{is}) \tag{1}$$

where $\lambda_{is} = \exp(x_i\beta_s + \varepsilon_{is})$. In MVPLN model, the latent effect ε_i defined for each observation is assumed to be uncorrelated with the explanatory variable x_i and follows a multivariate normal distribution with a mean vector of $\mathbf{0}$ and an unrestricted variance–covariance matrix, which is

$$\epsilon_i \Big| \Sigma \sim \varphi_s(0, \Sigma), \quad \text{for} \quad i = 1, 2, \dots, n \tag{2}$$

Where

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \cdots & \sigma_{1S} \\ \sigma_{21} & \sigma_{22} \cdots & \sigma_{2S} \\ \vdots & \vdots \ \vdots & \vdots \\ \sigma_{S1} & \sigma_{S2} \cdots & \sigma_{SS} \end{bmatrix} \tag{3}$$

and $\sigma_{sl} = \sigma_{ls}$, for $s, l \in \{1, 2, \dots, S\}$.

It should be noted that unlike the random parameter models (allowing parameter vary across observations) (Hensher and Greene, 2003; Milton et al., 2008) or finite mixture (latent class) methods (assume the sampled observations arise from distinct groups with homogeneous features) (Depaire et al., 2008; Eluru et al., 2012; Xiong and Mannering, 2013) that usually used in capture unobserved heterogeneities, the MVPLN model uses fixed parameters but introduces latent variable ε_i for each observation to accommodate the individual unobserved heterogeneity correlated among severity levels. As shown by Chib and Winkelmann (2001), the correlation between crash counts within segments can be positive or negative, which is unrestricted, depending on the sign of σ_{sl}. A positive σ_{sl} correspond to a positive correlation between $y_{is}$and y_{il}, vice versa. Moreover as σ_{ss} ($s = 1, \dots, S$) can also be positive, thus the model structure can also account for overdispersion.

3.2. Parameter estimation via MCMC simulation

As obtaining the marginal distribution for MVPLN model requires the evaluation of the S-variate integration of the Poisson distribution with respect to the distribution of ε_i, which is

$$P(y_i|\lambda_i, \Sigma) = \int \prod_{s=1}^{S} (y_{is}|\epsilon_i, \beta_s, x_{is}) \varphi_s(\epsilon_i|0, \Sigma) d\epsilon_i \tag{4}$$

Above distribution cannot be obtained through direct computation, thus the MCMC simulation approach is applied to estimate the unknown parameters under a Bayesian framework. For the prior distributions, it is assumed β and Σ independently follow a multivariate normal distribution and a Wishart distribution:

$$\beta \sim \varphi_k(\beta_0, V_{\beta 0}), \quad \Sigma^{-1} \sim f_W(v_\Sigma, V_\Sigma) \tag{5}$$

where $\phi_k(\cdot)$ is the probability density function of multivariate normal distribution with mean β_0 and covariance matrix V_{β_0}; $f_W(\cdot)$ is the Wishart distribution with degrees of freedom v_Σ and scale matrix V_Σ. The β_0, $V_{\beta 0}$, v_Σ and V_Σ are known hyperparameters. According to Bayes' theorem, the joint posterior density is proportional to:

$$posterior \propto prior \times likelihood = \phi_k(\beta_0, V_{\beta_0}) f_W(v_\Sigma, V_\Sigma) \prod_{i=1}^{n} \int \prod_{s=1}^{S} f_{poisson}(y_{is}|\varepsilon_i, \beta_s, x_{is}) \phi_s(\varepsilon_i|0, \Sigma) d\varepsilon_i \tag{6}$$

The joint posterior is then simulated by iteratively sampling from following three conditional posterior distributions:$\pi^P(\Sigma^{-1}|\varepsilon)$, $\pi^P(\varepsilon|y, X, \beta, \Sigma)$, and $\pi^P(\beta|y, X, \varepsilon, \Sigma)$. The sampling process is divided into three parts, which are sampling Σ^{-1}, ε and β accordingly.

3.2.1. Sampling Σ^{-1}

The posterior kernel of Σ^{-1} conditioned on data and other parameters can be written as

$$\pi^P(\Sigma^{-1}|\varepsilon) \propto f_W(\Sigma^{-1}|v_\Sigma, V_\Sigma) \prod_{i=1}^{n} \phi_s(\varepsilon_i|0, \Sigma) \tag{7}$$

By combining terms, the above posterior density is still a Wishart, which is

$$\Sigma^{-1}|\varepsilon \sim f_W\left(n + v_\Sigma, \left[V_\Sigma{}^{-1} + \sum_{i=1}^{n}\left(\varepsilon_i\varepsilon_i{}^T\right)\right]^{-1}\right) \tag{8}$$

Above distribution is a known parametric distribution and thus can be sampled using a Gibbs sampler.

3.2.2. Sampling ε in parallel

As the full posterior density for ε_i, $\pi^P(\varepsilon|y, X, \beta, \Sigma)$ is not given by any known density, thus the Metropolis–Hastings (M–H) algorithm is applied. The multivariate t distribution is used as the proposal density distribution, given as $f_T(\varepsilon_i|\hat{\varepsilon}_i, V_{\hat{\varepsilon}_i}, v_\varepsilon)$, where $\hat{\varepsilon}_i$ is obtained by maximizing the posterior probability for ε_i using Newton–Raphson algorithm:

$$\hat{\varepsilon}_i = \arg\max_{\varepsilon_i}\left[\ln \pi^P\left(\varepsilon_i|y_i, x_i, \beta, \Sigma\right)\right] \tag{9}$$

A proposal value $\varepsilon_i{}^*$ drawn from $f_T(\varepsilon_i|\hat{\varepsilon}_i, V_{\hat{\varepsilon}_i}, v_\varepsilon)$ is accepted with the following probability:

$$\alpha\left(\varepsilon_i, \varepsilon_i{}^*|y_i, x_i, \beta, \Sigma\right) = \min\left\{\frac{\pi^P\left(\varepsilon_i{}^*|y_i, x_i, \beta, \Sigma\right)f_T(\varepsilon_i|\hat{\varepsilon}_i, V_{\hat{\varepsilon}_i}, v_\varepsilon)}{\pi^P\left(\varepsilon_i|y_i, x_i, \beta, \Sigma\right)f_T(\varepsilon_i{}^*|\hat{\varepsilon}_i, V_{\hat{\varepsilon}_i}, v_\varepsilon)}, 1\right\} \tag{10}$$

Since the latent effect ε_i is independent between different observations, thus ε can be sampled in parallel for multiple observations. This scheme significantly speeds up the sampling process and shows superior performance especially when the number of observations is large.

3.2.3. Sample β in parallel

Similar to sampling ε_i, the M–H algorithm is used to sample β. As suggested by Chib and Winkelmann (2001), sampling β in one block may produce many rejections in M–H algorithm. An alternative approach is to sample the component of β, β_s one at a time. Again, the parallelization scheme is introduced to sampling multiple β_s's simultaneously. The multivariate t distribution $f_T(\beta_s|\hat{\beta}_s, V_{\hat{\beta}_s}, v_\beta)$ is again used as the proposal density distribution, with $\hat{\beta}_s$ obtained by maximizing the posterior probability for β_s using Newton–Raphson algorithm:

$$\hat{\beta}_s = \arg\max_{\beta_s}\left[\ln \pi^P\left(\beta_s|y_s, X, \varepsilon_s, \Sigma, \beta_{-s}\right)\right] \tag{11}$$

A proposal β_s^* is then drawn from $f_T(\beta_s|\hat{\beta}_s, V_{\hat{\beta}_s}, v_\beta)$ with the degree of freedom v_β. The Markov chain move to β_s^* from the current point β_s with a probability of

$$\alpha\left(\beta_s, \beta_s{}^*|y, x_i, \beta, \Sigma\right) = \min\left\{\frac{\pi^P\left(\beta_s^*|y_{.s}, X, \varepsilon_{.s}, \Sigma, \beta_{-s}\right)f_T\left(\beta_s|\hat{\beta}_s, V_{\hat{\beta}_s}, v_\beta\right)}{\pi^P\left(\beta_s|y_{.s}, X, \varepsilon_{.s}, \Sigma, \beta_{-s}\right)f_T\left(\beta_s^*|\hat{\beta}_s, V_{\hat{\beta}_s}, v_\beta\right)}, 1\right\} \tag{12}$$

3.3. Model comparison

The Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002) is a commonly used goodness of fit measure in Bayesian statistics. DIC is a Bayesian generalization of Akaike's Information Criteria (AIC) and Bayesian information criterion (BIC). Let θ be the parameter of the model, define $D(\theta) = -2\ln[p(y|\theta)] + 2\ln[f(y)]$ as Bayesian deviance, in which $p(y|\theta)$ is the likelihood function and $p(y)$ is a standardizing term that is a function of the data alone. Since $p(y)$ is a constant that cancels out in all calculations thus usually omitted in computation. The DIC is computed as

$$DIC = p_D + \overline{D}(\theta) = D(\bar{\theta}) + 2p_D \tag{14}$$

where $p_D = \overline{D}(\theta) - D(\bar{\theta})$. $\bar{\theta} = E[\theta|y]$ and $\overline{D}(\theta) = E[D(\theta)|y]$ are the posterior means of θ and the Bayesian deviance $D(\theta)$ respectively. A model with lower DIC indicates a superior model fit to the data. The DIC was used as the model comparison measure in the estimated MVPLN models in this paper. The final models were selected from the model with the set of explanatory variables that yields lowest DIC.

3.4. Elasticity of crash frequencies λ_{is}

To evaluate the relative impact of each variable in the model, the elasticity of the expected crash frequency λ_{is} is computed. Following the general formula for direct elasticity, the elasticity of the expected crash frequency λ_{is} for census tract i and severity level s is evaluated as follows:

$$E_{x_{ik}}^{\lambda_{is}} = \frac{\partial \lambda_{is}}{\lambda_{is}} \cdot \frac{x_{ik}}{\partial x_{ik}} = \beta_{sk} x_{ik} \tag{13}$$

Since the posterior means of the parameters $\bar{\beta}_s$, $\bar{\varepsilon}_{is}$ easily obtainable from the MCMC simulation, we compute the elasticity of frequency λ_{is} evaluated at the posterior means of the parameters ($\bar{\beta}_s$, $\bar{\varepsilon}_{is}$), and the average elasticity overall all census tracts is reported.

4. Data

To fully demonstrate the analytical capability of the MVPLN model, we have used two datasets: (a) census level pedestrian–vehicle crash data with two severity levels (Shin et al., 2010; New York City Department of Transportation (NYCDOT), 2007) from New York City (5 year data – from 2002 to 2006), and (b) highway crash data (Milton et al., 2008) with three severity levels from Washington State (5 year data – from 1990 to 1994).

The pedestrian–vehicle crash dataset serves as the primary dataset for the analysis in this paper, since it is more comprehensive with 2183 observations (census tracts of New York City) and 90 explanatory variables. The dataset was compiled by Center for Transportation Injury Research at CUBRC from multiple sources and used by several studies to analyze pedestrian–vehicle crashes in New York City (Ukkusuri et al. 2011; Aziz et al., 2013; Mohamed et al., 2013; Yasmin et al., 2014). However, these studies either focused on modeling crash severity levels or use univariate count data model by aggregating pedestrian–vehicle crashes across severity levels. The dataset contains 637 fatal pedestrian–vehicle crashes and 5790 severe injury crashes. In the dataset, for each of 2183 census tracts of New York City, the numbers of fatal and severe pedestrian–vehicle crashes were aggregated and the corresponding demographic information, land use patterns and traffic system characteristics were also gathered. Table 1 gives the descriptive statistics of the 16 explanatory variables used in the developed model.

One limitation of the pedestrian–vehicle crash dataset is that it only contains two severity levels: fatal and severe injury crashes, which is insufficient to fully illustrate the potential of the MVPLN model. Consequently, a second MVPLN model using the Washington State highway injury dataset with three severity levels (no-injury, possible injury, evident injury) was developed to demonstrate the applicability of the proposed estimation procedure and tool on multivariate cases. The dataset is from a previous study by Milton et al. (2008), which consist of crash injury data of 275 roadway segment from 1990 to 1994 in Washington State with 30 explanatory variables. Detailed information of this dataset please refer to Milton et al. (2008). Since the dataset is much smaller and less detail compared to the pedestrian–vehicle crash dataset, it is used as a supplement. Table 2 provides the descriptive statistics of the 9 explanatory variables used in the supplementary MVPLN model.

Table 1
Descriptive statistics of selected variables (NYC pedestrian–vehicle crash dataset).

Variable description	Mean	Std. dev.	Minimum	Maximum
Dependent variable				
Number of fatal crashes	0.292	0.613	0	5
Number of severe injury crashes	2.652	3.197	0	34
Demographic characteristics				
Tract population in 2000 (in 10,000)	0.367	0.244	0.003	2.452
Black population proportion	0.281	0. 322	0	0.980
Population aged 65 and over proportion	0.119	0.066	0	0.900
Population aged 25 years and over with high school education proportion	0.326	0.101	0	0.826
Land use attributes				
Industrial/manufacturing land use proportion	0.039	0.066	0	0.605
Number of schools	1.077	1.410	0	13
Road network and intersection operation characteristics				
Signalized intersections (in 10)	0.561	0.474	0	10.5
Length of roads in miles in tract (in 10 miles)	0.384	0.566	0.02	10.421
Primary roads without limited access proportion of total roadway length	0.028	0.080	0	0.993
Local, neighborhood, and rural roads proportion of total roadway length	0.702	0.207	0	0.999
Roads in tract with 5 or more travel lanes proportion of total roadway length	0.050	0.023	0	0.238
Roads with lane widths less than 30 ft proportion of total roadway length	0.233	0.208	0	0.993
Truck routes proportion of total roadway length	0.166	0.175	0	0.985
Subway stations in tract	0.224	0.541	0	7

Table 2
Descriptive statistics of selected variables (Washington State highway injury dataset).

Variable description	Mean	Std. dev.	Minimum	Maximum
Dependent variable				
Number of no-injury crashes	9.749	12.587	0	87
Number of possible injury crashes	3.415	5.623	0	32
Number of evident injury crashes	3.771	4.762	0	35
Explanatory variables				
Highway segment length in miles	2.430	2.694	0.5	19.3
Logarithm of average annual daily travel per lane	8.631	0.763	6.730	10.264
Maximum grade difference in the segment	3.039	1.983	0	9.8
Number of horizontal curves per mile in the segment	1.440	0.961	0	5
Percentage of trucks (all truck types) in the traffic	14.163	6.692	3.2	32
Low precipitation indicator (≤12 in. per year)	0.324	0.469	0	1
Heavy snow fall indicator (≥18 in. per year)	0.178	0.383	0	1
Local road indicator	0.305	0.461	0	1

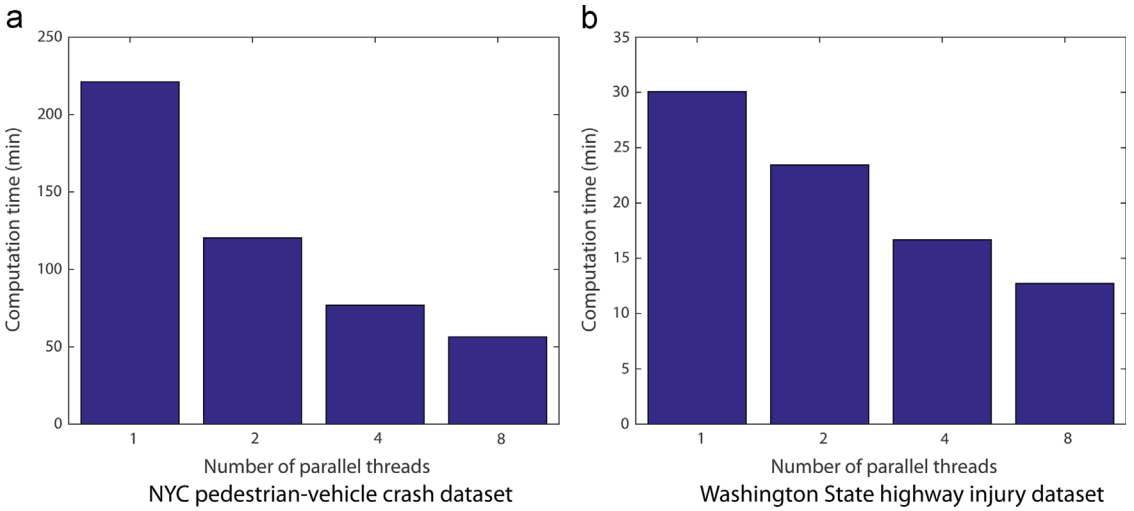

Fig. 1. Computation performance of performing 10,000 MCMC sample draws on two datasets.

5. Results and discussions

A new tool was developed in MATLAB to estimate the MVPLN model implementing the parallelized MCMC sampling scheme described in the methodology. For the estimation of the MVPLN model, 10,000 MCMC sample draws were used and the first 4000 sample draws were discarded. The trace plots for all of the model parameters were inspected to ensure the convergence. A second chain using different initial values for each model was also performed to verify the convergence of the estimates to the same set of posterior means. Fig. 1 shows the computation performance of the parallel sampling scheme by performing 10,000 MCMC samples draws on the two datasets studied in this paper. All tests were conducted on an Intel i7 2.3 GHz CPU laptop. The computation time can be further reduced if more parallel threads were used. The parallelized sampling scheme effectively reduced the computation time. For the larger NYC pedestrian–vehicle crash dataset (2 severities, 2188 observations), the estimation process without parallelization takes 221 min while with 8 parallel threads, the computation time is reduced to 56 min. For the smaller Washington State highway injury dataset (3 severities, 275 observations), with 8 parallel threads, the entire estimation process just takes less than 13 min.

As the parameters were estimated using a MCMC approach, the high density region (HDR) containing 95% (2.5–97.5%) of the sample coefficient values was used to inspect whether a parameter is statistically significant. The criteria for variable selection in developing the two models in this paper are: (1) the variable was statistically significant in all severity levels or had significant impact on one severity level; (2) the inclusion of the variable improved the model goodness-of-fit measure.

For the MVPLN model developed using NYC pedestrian–vehicle crash dataset, the estimated coefficients are presented in Table 3 for both fatal and severe injury crashes. Fig. 2 plots the probability density distribution from the sampled parameter values, which provides a more intuitive representation of the differences between sampled parameters on the two severity levels. The fatal crashes have a “fatter” distribution, mainly because of fewer observations for fatal crashes, thus lead to larger variance in the sampled coefficients. Table 4 presents the estimation results of covariance matrix (Σ) of latent effects.

Table 3
Estimation result of MVPLN model (NYC pedestrian–vehicle crash dataset, underscore marks the variable that is not significant in the severity level).

Variable description	Fatal crashes					Severe injury crashes				
	Mean	Std.	The 95% HDR		Elasticity	Mean	Std.	The 95% HDR		Elasticity
Constant	−2.152	0.205	−2.573	−1.754		0.141	0.083	−0.024	0.308	
Demographic characteristics										
Tract population in 2000	0.844	0.146	0.542	1.132	0.310	1.083	0.057	0.968	1.199	0.397
Black population proportion	−0.376	0.155	−0.688	−0.078	−0.106	0.163	0.056	0.053	0.275	0.046
Population age 65 and over proportion	1.247	0.614	0.038	2.462	0.148	−1.464	0.305	−2.073	−0.846	−0.174
Population aged 25 years and over with high school education proportion	0.971	0.395	0.188	1.770	0.317	0.496	0.156	0.189	0.808	0.162
Land use attributes										
Industrial/manufacturing land use proportion	1.864	0.554	0.701	2.935	0.073	3.071	0.201	2.671	3.475	0.120
Number of schools	0.056	0.026	0.004	0.107	0.060*	0.038	0.010	0.018	0.059	0.041*
Road network and intersection operation characteristics										
Signalized intersections	0.486	0.074	0.340	0.636	0.273	0.572	0.032	0.508	0.636	0.321
Length of roads in miles in tract (in 10 miles)	−0.225	0.091	−0.4140	−0.055	−0.086	−0.386	0.044	−0.476	−0.300	−0.148
Primary roads without limited access proportion	0.483	0.302	−0.123	1.055	0.015	0.338	0.129	0.075	0.587	0.010
Local, neighborhood, and rural roads proportion	−0.421	0.091	−0.600	−0.238	−0.460	−0.259	0.033	−0.326	−0.195	−0.282
Roads with 5 or more travel lanes proportion	3.169	1.258	0.604	5.609	0.016	2.036	0.541	0.949	3.110	0.010
Roads with lane widths less than 30 ft proportion	−0.467	0.223	−0.909	−0.028	−0.109	−0.470	0.089	−0.649	−0.295	−0.109
Truck routes proportion	0.789	0.188	0.401	1.169	0.142	0.552	0.076	0.403	0.700	0.099
Subway stations in tract	0.053	0.067	−0.085	0.186	0.012*	0.107	0.025	0.055	0.157	0.024*

Summary statistics				
Number of observations	$LL(\bar{\theta})$	$\bar{D}$	p_D	DIC
2183	−5298.72	14,785.91	4188.48	18,974.40

Note: 1. High density region (HDR) is used to inspect whether the parameters differ from zero in a statistically significant way, which is computed as the region that contains 95% (2.5–97.5%) sampled parameter values.
2. The elasticity reported in the table is the average elasticity of frequency λ evaluated at the posterior mean of the model parameters.
3. As the number of schools and subway stations are discrete integer valued variables, the elasticity is not directly applicable. Hence we just report the computed value (marked in *). However, as the tract population is in 10,000, and the signalized intersections are in 10, thus we treat these variables as continuous.

Univariate Poisson Lognormal (PLN), univariate Poisson and univariate Negative Binomial (NB) models were estimated for the same set of explanatory variables (see Table 7). The same set of results for the MVPLN model that developed using the Washington State highway injury dataset are also reported in Fig. 3 and Tables 5, 6 and 8. The following subsections briefly discuss the goodness-of-fit characteristics, correlation among severity levels and parameter estimates.

5.1. Model fitting and comparison

The DIC was used as the model comparison measure in developing the MVPLN model. The final model has the specification with the lowest DIC. Tables 3 and 5 provide the DIC value for the two MVPLN models, while the DIC values for the corresponding univariate PLN models can be found in Tables 7 and 8. Furthermore, the log-likelihood evaluated at the posterior means of the parameters of the MVPLN model and the two univariate PLN models are computed. The log-likelihood at convergence of univariate Poisson and Negative Binomial are also reported, which serve as a comparison purpose.

For the NYC pedestrian–vehicle crash model, compared with the DIC of 6616.84 under fatal PLN model and 12,526.14 under severe injury PLN model, the DIC statistics for the MVPLN model is only 18,974.40. A significant drop of 168.58 is observed compared with the sum of the two univariate models. For the Washington State highway injury model, an even large drop of 799.59 in DIC value is observed for the MVPLN model (4852.18) compared with three univariate PLN counterparts (2138.60, 1748.47 and 1764.70 respectively). The larger drop in DIC is likely related to higher level of correlation among severity levels in Washington State highway injury dataset (see Table 6). The DIC statistics of the MVPLN model for both of the pedestrian–vehicle and highway-injury dataset clearly indicates that MVPLN provides a much better fit compared with the univariate PLN models that ignore the underlying correlation among severity levels. Furthermore, it is observed that when the correlation among severity levels are high, the estimated coefficients in MVPLN model have higher

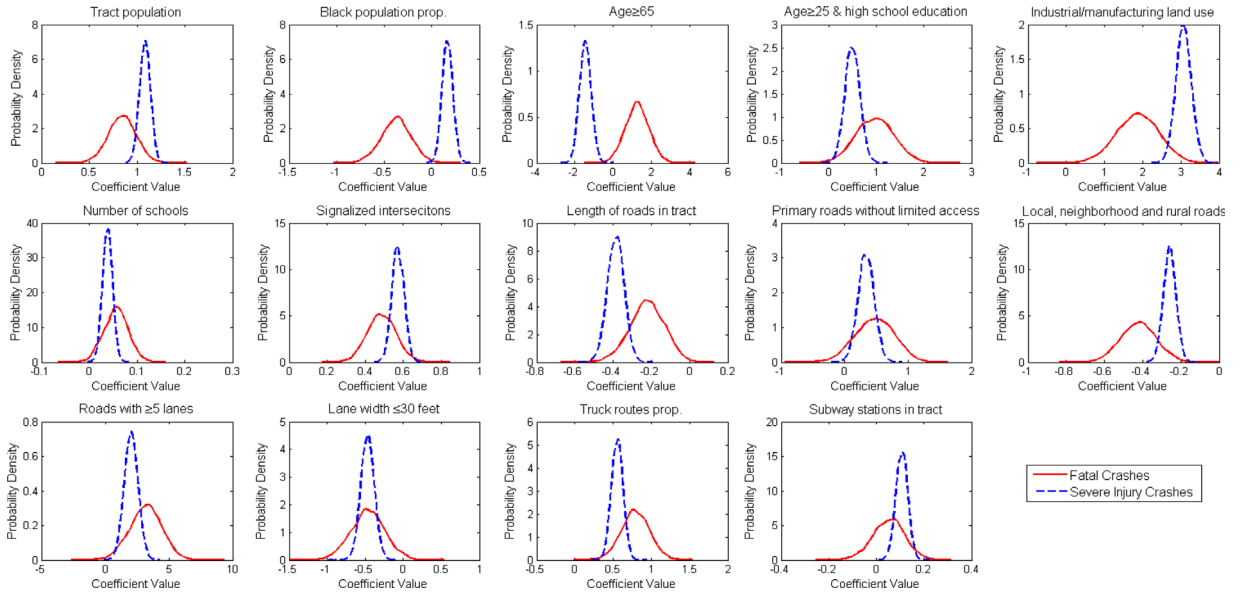

Fig. 2. Probability density distribution of the sampled parameters (NYC pedestrian–vehicle crash dataset).

Table 4
Estimation result of covariance matrix (Σ) of latent effects (NYC pedestrian–vehicle crash dataset).

σ_{ij}	Mean	Std. Err.	The 95% HDR	
σ_{11} (fatal)	0.270	0.051	0.179	0.383
σ_{12}, σ_{21}	0.185	0.025	0.138	0.237
σ_{22} (severe injury)	0.253	0.020	0.215	0.296
Correlation $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$	0.710	0.058	0.581	0.812

precision than univariate PLN models. This can be shown in the model comparison results of the Washington State highway injury dataset in Table 8, that the estimated standard deviation of coefficients for MVPLN model are always lower than the corresponding PLN model.

By comparing the log-likelihood evaluated at the posterior mean of parameters, the MVPLN model also outperforms the univariate PLN, Poisson and Negative Binomial models. For the NYC pedestrian–vehicle crash dataset, the MVPLN model provides a 79.4 greater of log-likelihood value than the sum of the two univariate PLN models, 513.9 and 209.5 greater than the sum of the two univariate Poisson and Negative Binomial models. For the Washington State highway injury dataset, the MVPLN provides a 400.21 greater of log-likelihood value than the sum of the three univariate PLN models, 989.75 and 321.38 greater than the sum of the three univariate Poisson and Negative Binomial models. All of the above statistics demonstrate the superiority of MVPLN in modeling crash data with multiple correlated severity levels, in which cases using univariate models would lead to less accurate parameter estimates and even biased policy recommendations.

5.2. Correlation among latent effects across severity levels

The correlation values of latent effects among different severity levels were also quantified for the two dataset investigated. This correlation is a measure of dependencies for the latent effects of different severity levels, rather than the correlation directly obtained from crash counts, which is trivial to compute. High level of correlations among severity levels are observed in this study. For the correlation among fatal and severe injury pedestrian–vehicle crashes at census tract level, the correlation value of the latent effects is found to be 0.710, and the 95% high density region (0.5812, 0.8122) shows that the estimated correlation is highly significant (see Table 4). The level of correlation in latent effects between fatal and severe injury pedestrian–vehicle crashes in this paper is found similar with the correlation values for vehicle crashes in the literature. Park and Lord (2007) found that the correlation between fatal and incapacitating-injury crashes was 0.7035, and correlation between fatal and non-incapacitating injury crashes was 0.6904. A study by Aguero-Valverde and Jovanis (2009) reported the correlation between fatal and incapacitating-injury crashes was 0.542, and the correlation between fatal and

Table 5
Estimation results of MVPLN model (Washington State highway injury dataset, underscore marks the variable that is not significant in the severity level).

Variable description	No-injury crashes					Possible injury crashes					Evident Injury Crashes				
	Mean	Std.	The 95% HDR		Elasticity	Mean	Std.	The 95% HDR		Elasticity	Mean	Std.	The 95% HDR		Elasticity
Constant	−2.744	0.639	−3.991	−1.475		−5.768	1.041	−7.789	−3.716		−1.864	0.915	−3.686	−0.068	
Explanatory Variables															
Highway segment length in miles	0.140	0.009	0.123	0.157	0.341	0.145	0.015	0.115	0.176	0.353	0.161	0.011	0.140	0.182	0.390
Logarithm of AADT per lane	0.592	0.061	0.472	0.713	5.108	0.856	0.099	0.654	1.053	7.392	0.368	0.088	0.194	0.546	3.179
Maximum grade difference	0.062	0.015	0.033	0.093	0.190	0.030	0.025	−0.019	0.078	0.091	0.105	0.022	0.063	0.149	0.319
Number of horizontal curves per mile	−0.174	0.036	−0.246	−0.103	−0.250	−0.251	0.055	−0.361	−0.139	−0.362	−0.249	0.053	−0.357	−0.145	−0.359
Percentage of trucks in traffic	−0.027	0.007	−0.041	−0.014	−0.385	−0.054	0.011	−0.076	−0.032	−0.765	−0.028	0.009	−0.046	−0.009	−0.393
Low precipitation indicator	−0.501	0.093	−0.690	−0.315	–	−0.667	0.164	−0.996	−0.349	–	−0.343	0.132	−0.614	−0.083	–
Heavy snow fall indicator	−0.212	0.083	−0.377	−0.041	–	−0.164	0.136	−0.426	0.106	–	−0.254	0.125	−0.509	−0.008	–
Local road indicator	−0.338	0.077	−0.492	−0.187	–	−0.188	0.118	−0.428	0.044	–	−0.293	0.116	−0.521	−0.057	–
Summary statistics															
Number of Observations	$LL(\bar{\theta})$		$\bar{D}$			p_D					DIC				
275	−1636.503		4062.59			789.59					4852.18				

Note: The last four variables are binary indicator variables, thus the elasticity is not directly applicable and not presented.

Table 6
Estimation results of covariance matrix (Σ) of latent effects (Washington State highway injury dataset).

	Mean	Std. Err.	The 95% HDR	
σ_{11} (No-injury)	0.663	0.086	0.508	0.855
σ_{12}, σ_{21}	0.657	0.090	0.498	0.859
σ_{13}, σ_{31}	0.555	0.077	0.417	0.722
σ_{22} (Possible injury)	0.657	0.090	0.498	0.859
σ_{23}, σ_{32}	0.797	0.128	0.571	1.091
σ_{33} (Evident injury)	0.588	0.095	0.419	0.792
Correlation				
$\rho_{12} = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$	0.906	0.023	0.853	0.944
$\rho_{13} = \sigma_{13}/\sqrt{\sigma_{11}\sigma_{33}}$	0.891	0.028	0.827	0.938
$\rho_{23} = \sigma_{23}/\sqrt{\sigma_{22}\sigma_{33}}$	0.868	0.036	0.784	0.929

Table 7
Estimated coefficients comparison of MVPLN, Univariate Poisson-lognormal (PLN), Univariate Poisson, and Negative Binomial (NB) Model (NYC pedestrian–vehicle crash dataset, values in each cell are estimated coefficient and the standard deviation in parentheses).

Variable description	MVPLN	PLN	Poisson	NB
Fatal crashes				
Constant	−2.152(0.205)	−2.059(0.205)	−1.974(0.193)	−2.078(0.232)
Tract population in 2000	0.844(0.146)	0.800(0.139)	0.790(0.129)	0.776(0.163)
Black population proportion	−0.376(0.155)	−0.383(0.156)	−0.388(0.151)	−0.364(0.160)
Population age 65 and over proportion	1.247(0.614)	1.263(0.605)	1.281(0.590)	1.274(0.708)
Population aged 25 years and over with high school education proportion	0.971(0.395)	0.948(0.383)	0.919(0.384)	0.995(0.451)
Industrial/manufacturing land use proportion	1.864(0.554)	1.771(0.518)	1.738(0.514)	1.752(0.572)
Number of schools	0.056(0.026)	0.054(0.026)	0.054(0.025)	0.053(0.029)
Signalized intersections	0.486(0.074)	0.468(0.075)	0.456(0.073)	0.506(0.069)
Length of roads in miles in tract (in 10 miles)	−0.225(0.091)	−0.232(0.099)	−0.239 (0.096)	−0.186 (0.081)
Primary roads without limited access proportion	0.483(0.302)	0.464(0.303)	0.526(0.280)	0.514(0.340)
Local, neighborhood, and rural roads proportion	−0.421(0.091)	−0.437(0.095)	−0.432(0.090)	−0.400(0.078)
Roads with 5 or more travel lanes proportion	3.169(1.258)	3.033(1.263)	3.033(1.208)	3.236(1.548)
Roads with lane widths less than 30 feet proportion	−0.467(0.223)	−0.443(0.219)	−0.408(−0.218)	−0.438(0.237)
Truck routes proportion	0.789(0.187)	0.761(0.183)	0.767(0.180)	0.760(0.214)
Subway stations in tract	0.053(0.067)	0.053(0.064)	0.052(0.061)	0.061(0.062)
Log-likelihood ($LL(\bar{\theta})$ for Univariate PLN)	–	−1224.09	−1394.03	−1388.63
DIC	–	6616.84	Dispersion: 0.328	
Severe Injury Crashes				
Constant	0.141(0.083)	0.138(0.082)	0.326(0.067)	−0.166(0.101)
Tract population in 2000	1.083(0.057)	1.083(0.058)	0.996(0.042)	1.083(0.078)
Black population proportion	0.163(0.056)	0.164(0.056)	0.138(0.047)	0.201(0.071)
Population age 65 and over proportion	−1.464(0.305)	−1.423(0.298)	−1.345(0.250)	−1.445(0.363)
Population aged 25 years and over with high school education proportion	0.496(0.156)	0.490(0.152)	0.525(0.129)	0.648(0.200)
Industrial/manufacturing land use proportion	3.071(0.201)	3.063(0.202)	2.853(0.155)	3.072(0.314)
Number of schools	0.038(0.010)	0.038(0.010)	0.033(0.008)	0.032(0.015)
Signalized intersections	0.572(0.032)	0.571(0.032)	0.551(0.026)	0.803(0.039)
Length of roads in miles in tract (in 10 miles)	−0.386(0.044)	−0.388(0.043)	−0.441(0.038)	−0.223(0.020)
Primary roads without limited access proportion	0.338(0.129)	0.336(0.128)	0.353(0.099)	0.214(0.191)
Local, neighborhood, and rural roads proportion	−0.259(0.033)	−0.260(0.033)	−0.281(0.028)	−0.164(0.019)
Roads with 5 or more travel lanes proportion	2.036(0.541)	2.019(0.536)	1.670(0.434)	2.430(0.790)
Roads with lane widths less than 30 feet proportion	−0.470(0.089)	−0.465(0.089)	−0.459(0.074)	−0.368(0.103)
Truck routes proportion	0.552(0.076)	0.555(0.074)	0.522(0.062)	0.513(0.105)
Subway stations in tract	0.106(0.025)	0.106(0.025)	0.090(0.019)	0.127(0.034)
Log-likelihood ($LL(\bar{\theta})$ for Univariate PLN)	–	−4154.03	−4418.547	−4119.58
DIC	–	12,526.14	Dispersion: 0.343	

non-incapacitating injury was 0.584. El-Basyouny and Sayed (2009) reported the correlation between no-injury crashes and injuries and fatal (I+F) crashes was 0.758.

For vehicle crashes on highway segments of Washington State, the correlation values of the latent effects were estimated to be 0.906 between the no-injury and possible injury crashes; 0.891 between no-injury and evident injury crashes; and

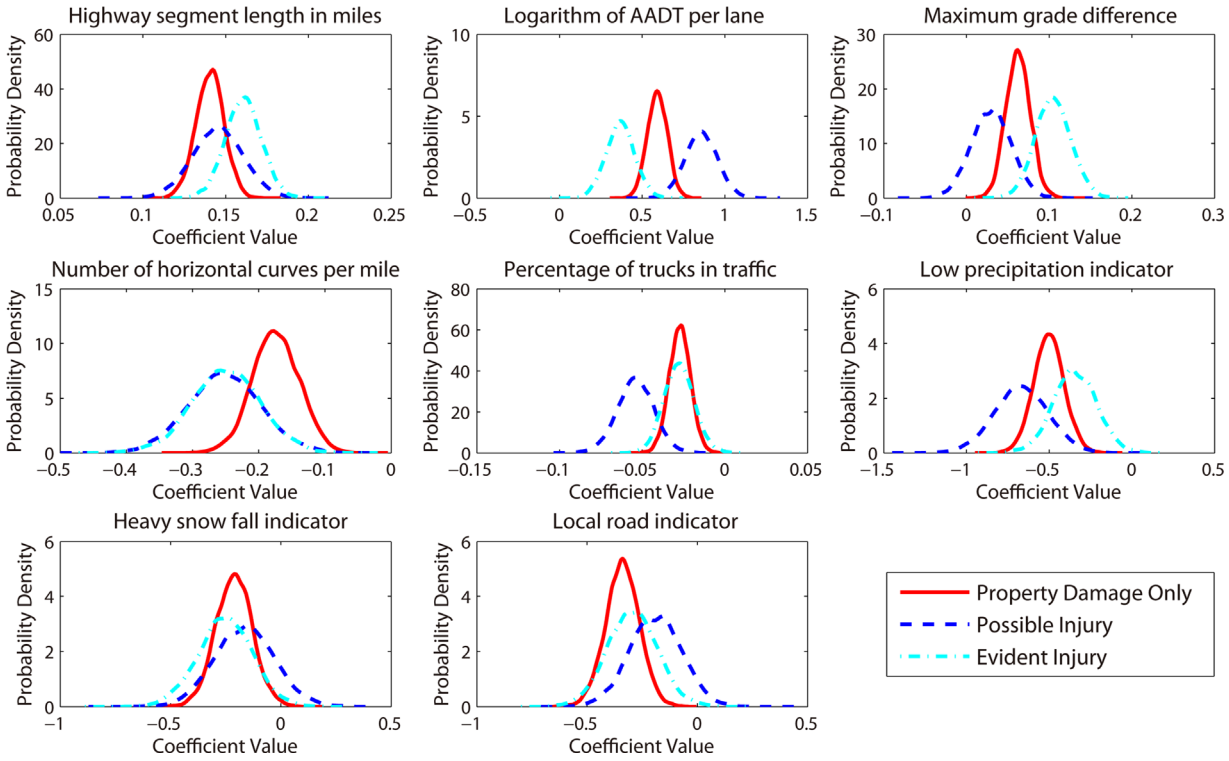

Fig. 3. Probability density distribution of the sampled parameters (Washington State highway injury dataset).

0.868 between possible and evident injury crashes (see Table 6). All of the estimated correlation values are found to be highly significant. The reason for the higher correlation level in the highway injury data compared with pedestrian–vehicle crash data might due to the relatively vague boundary among the three injury severity levels. The correlation results confirm the intuition that correlations exist among severity levels for both pedestrian–vehicle crashes and highway vehicle crashes. In addition, it suggests the necessity of jointly modeling crashes at multiple severity levels. Univariate models that ignore the underlying correlations among severity levels may lead to misspecification of models and inappropriate result interpretations.

5.3. Parameter estimates

5.3.1. NYC pedestrian–vehicle crash dataset

Table 3 and Fig. 2 summarize the coefficient estimates and corresponding statistics for the MVPLN model using NYC pedestrian–vehicle crash dataset. The results indicate that, the census tracts with higher population are more likely to have higher crash frequency both at fatal and severe levels of injuries. Since we do not have the traffic volume variable in our model, population is an indirect measure of number of pedestrians and traffic volume in our study. Previous studies (Narayanamoorthy et al., 2013; Ukkusuri et al., 2011; LaScala et al., 2000) also found population density as a significant variable. The average elasticity of the expected crash frequency suggests that 1% increase in 10,000 population increases the expected frequencies of fatal crashes by 0.31% and severe injury crashes by 0.397%. The proportion of black population is often considered as an indicator of low-income population in different regions of the United States. Based on previous studies it is possible to correlate low-income neighborhoods with lower level of educated populations, and higher number of pedestrians. In our study, the proportion of black population is found to have very different effects on fatal and severe injury crash frequencies, and clear difference can be observed in the probability density plot (Fig. 2). Results show that, higher proportion of black population is expected to increase the frequency of severe injury crashes, while decrease the frequency of fatal crashes. The average elasticity suggests that 1% increase in the black population proportion will decrease the fatal crash frequency by 0.106%, but will increase the expected severe injury crash frequency by 0.046%. The decrease in likelihood of fatal crashes is likely to be associated with lower driving speed in such neighborhood with poorer condition of transportation facilities. Similarly, census tracts with higher proportion of population with high school level education are found more prone to pedestrian vehicle crashed at both level of severities.

Further, census tracts with higher proportion of elderly population (age 65 years or above) are more prone to fatal crashes. Again completely different parameter probability density distribution is observed, that the distribution for fatal crashes peaks around 1.5 while for severe injury crashes, peaks around −1.5. This is intuitive because most crashes

Table 8
Estimated Coefficients Comparison of MVPLN, Univariate Poisson-lognormal (PLN), Univariate Poisson, and Negative Binomial (NB) Model (Washington State highway injury dataset, values in each cell are estimated coefficient and the standard deviation in parentheses).

Variable Description	MVPLN	PLN	Poisson	NB
No-injury crashes				
Constant	−2.744(0.639)	−2.885(0.666)	−2.699(0482)	−3.798 (1.268)
Highway segment length in miles	0.140(0.009)	0.144(0.009)	0.137(0.006)	0.185(0.024)
Logarithm of AADT per lane	0.592(0.061)	0.603(0.063)	0.590(0.046)	0.677(0.123)
Maximum grade difference	0.062(0.015)	0.064(0.015)	0.059(0.011)	0.075(0.034)
Number of horizontal curves per mile	−0.174(0.036)	−0.174(0.037)	−0.168(0.025)	−0.135(0.068)
Percentage of trucks in traffic	−0.027(0.007)	−0.027(0.007)	−0.027(0.005)	−0.023(0.014)
Low precipitation indicator	−0.501(0.093)	−0.501(0.095)	−0.463(0.072)	−0.433(0.178)
Heavy snow fall indicator	−0.212(0.083)	−0.205(0.086)	−0.192(0.063)	−0.111(0.199)
Local road indicator	−0.338(0.077)	−0.329(0.080)	−0.327(0.057)	−0.274(0.168)
Log-likelihood ($LL(\bar{\theta})$ for Univariate PLN)	–	−804.42	−1251.78	−821.04
DIC	–	2138.60	Dispersion: 0.522	
Possible injury crashes				
Constant	−5.768(1.041)	−5.972(1.108)	−5.533(0.858)	−6.499(1.939)
Highway segment length in miles	0.145(0.015)	0.157(0.016)	0.141(0.011)	0.195(0.031)
Logarithm of AADT per lane	0.856(0.099)	0.872(0.105)	0.835(0.082)	0.914(0.188)
Maximum grade difference	0.030(0.025)	0.035(0.025)	0.032(0.018)	0.038(0.044)
Number of horizontal curves per mile	−0.251(0.055)	−0.253(0.056)	−0.258(0.043)	−0.189(0.089)
Percentage of trucks in traffic	−0.054(0.011)	−0.055(0.012)	−0.052(0.009)	−0.056(0.019)
Low precipitation indicator	−0.667(0.164)	−0.699(0.174)	−0.637(0.139)	−0.598(0.261)
Heavy snow fall indicator	−0.164(0.136)	−0.198(0.145)	−0.196(0.112)	−0.023(0.245)
Local road indicator	−0.188(0.118)	−0.184(0.125)	−0.196(0.093)	−0.135(0.250)
Log-likelihood ($LL(\bar{\theta})$ for Univariate PLN)	–	−614.22	−681.73	−538.17
DIC	–	1748.47	Dispersion: 0.652	
Evident injury crashes				
Constant	−1.864(0.915)	−2.047(0.951)	−1.818(0.738)	−2.611(1.576)
Highway segment length in miles	0.161(0.011)	0.164(0.012)	0.152(0.008)	0.195(0.026)
Logarithm of AADT per lane	0.368(0.088)	0.378(0.092)	0.368(0.072)	0.428(0.149)
Maximum grade difference	0.105(0.022)	0.108(0.022)	0.098(0.017)	0.114(0.036)
Number of horizontal curves per mile	−0.249(0.053)	−0.241(0.053)	−0.235(0.042)	−0.212(0.079)
Percentage of trucks in traffic	−0.028(0.009)	−0.027(0.010)	−0.026(0.008)	−0.026(0.016)
Low precipitation indicator	−0.343(0.132)	−0.358(0.140)	−0.347(0.109)	−0.308(0.227)
Heavy snow fall indicator	−0.254(0.125)	−0.222(0.131)	−0.240(0.101)	−0.063(0.194)
Local road indicator	−0.293(0.116)	−0.296(0.118)	−0.306(0.091)	−0.251(0.185)
Log-likelihood ($LL(\bar{\theta})$ for Univariate PLN)	–	−620.07	−692.74	−598.67
DIC	–	1764.70	Dispersion: 0.466	

involving elderly individuals will have higher likelihood to result in fatal crashes. It is observed that census tracts with higher number of schools are more vulnerable to pedestrian–vehicle crashes, and the impact is found to be similar for both severity levels (probability density distribution overlaps). The studies by Loukaitou-Sideris et al. (2007) and Narayanamoorthy et al. (2013) had similar results with the explanation that number of schools can be an indirect exposure related positive effect for the crash frequency in the census tract.

For land use related attributes, it is found that census tracts with higher proportion of industrial or manufacturing land use are more likely to have more pedestrian–vehicle crashes. The results show the same trends for fatal and severe injury crashes, and greater impact is found for severe injury crashes (Fig. 2 and higher elasticity value). Similar findings were also reported by Kim and Yamashita (2002) and Ukkusuri et al. (2011) and Narayanamoorthy et al. (2013). The reason could be associated with higher volume of truck traffic. Trucks often have larger blind spot and are more likely to result in more severe crashes compared with other vehicle types. This is confirmed in the result of proportion of truck routes variable, in which its average elasticity suggests that 1% increase in the truck routes proportion will increase the expected frequency of fatal crashes by 0.142% and severe injury crashes by 0.099%.

Very similar impact of road geometry related factors, such as presence of roadways with five or more lanes and lane width smaller or equal to 30 ft are found in both severity levels, as their probability density distributions almost overlaps, especially for the lane width indicator (Fig. 2). This suggests the impact of road geometry factors do not differentiate much between severity levels. This may be because that such kind of pedestrian–vehicle crashes usually occurs when pedestrian crosses the road, and the level of severity is largely depend on the specific situation of the crash rather than the road geometries. Roadways with five or more lanes increase the frequency of pedestrian–vehicle crashes, while narrow roadways

decrease the frequency. This is intuitive, since more lanes increase the chance of pedestrian–vehicle conflict, while narrow road usually has lower driving speed and cautious drivers. Primary roads without limited access is also found to increase the pedestrian–vehicle crash frequency and have very similar impact on both of the severity levels. This is intuitive since primary roads usually have higher travel speed, thus if pedestrian–vehicle conflicts exit in these type of roadways (e.g. without limited access), it would be more likely to have crashes. Local, neighborhood, and rural roads on the other hand, are found to have lower likelihood to have pedestrian–vehicle crashes. However, the probability density plot shows the fatal crash frequency is decreased more than severe injury crashes. The average elasticity shows that 1% increase in the proportion of these types of roads will decrease the expected frequency of fatal crashes by 0.46% and severe injury crashes by 0.282%. This may be largely related with the lower traffic volume and vehicle speed on local, neighborhood and rural roads, which leads to fewer crashes and lower severity levels.

The number of subway stations in the census tracts is found to increase the probability of severe injury pedestrian–vehicles crashes in census tracts, and the impact on fatal crashes is little. Higher number of subway stations can be an indirect measure of the pedestrian activities, which may result in higher chance of pedestrian vehicle conflict. However, since these census tracts are also most congested areas in the city, thus the chance to have fatal crashes is low.

5.3.2. Washington State highway injury dataset

The model estimation results for the supplementary Washington State highway injury dataset are summarized in Table 5 and Fig. 3. There are 8 explanatory variables used in the model, all of which have significant coefficients in all severity levels. The highway segment length is found to increase the frequency of crashes on all severity levels, which is intuitive. The logarithm of AADT per lane is found to be positively related to crashes of all severity levels. Similar finding was also reported in Barua et al. (2014). Higher volume of traffic is observed to have larger impact on possible injury crashes, as the elasticity for logarithm of AADT per lane is 7.392 for possible injury crashes, while the number is 5.108 for no-injury crashes and 3.179 for evident injury crashes. A large grade difference in highway segment is potentially causing disruption in normal driving behavior, which is also reflected in the model estimates. The coefficients for maximum grade difference in segment are positive in all severity levels. It has a particularly large impact on evident injury crashes with the highest coefficient value of 0.105 and 1% increase in grade difference in segment increase the expected evident injury crash frequency by 0.319%, which is higher than the other two severity levels.

Factors related to decreased driving speed such as number of horizontal curves per mile, local road indicator are again found to have negative coefficient that decrease crash frequencies in all severity levels. An interesting finding is that the percentage of truck in traffic is found to decrease the crash frequencies in all severity levels on highways, whereas it is found to increase crash frequencies for pedestrian–vehicle crashes in urban areas. The reason might still be related to the lower driving speed in these cases on highways. Since trucks usually travel slower than passenger vehicles, higher proportion of truck traffic would potentially decrease the overall traveling speed of vehicle on the highway segment and lead to more cautious driving behaviors. Highway segments with lower level of precipitation ($\leq$12 in. per year) are found negatively related to crashes of all the three injury levels. This might associate with better roadway friction condition. Highway segments with higher level of snow fall ($\geq$18 in. per year) are also found to have lower number of crashes in all injury levels. The impacts are almost the same for the three injury levels as their probability density of the sampled parameters almost overlaps (see Fig. 3). The negative coefficient values for heavy snow fall indicator are likely to pick up the geographical differences in driving behaviors in response to snow. Since drivers tend to have more cautious driving behavior in response to heavy snow fall. Similar negative coefficient estimates were also observed in Milton et al. (2008).

6. Concluding remarks

This study investigated the potential of using Multivariate Poisson-lognormal (MVPLN) model in multi-severity level crash data analysis. Two datasets were used to fully demonstrate the analytical capability of the MVPLN model in different applications. The primary dataset used is a comprehensive pedestrian–vehicle crash dataset which has both number of fatal and severe injury crashes from 2183 census tracts in New York City. Variables involving demographic characteristics, land use attributes and road network and intersection operation characteristics are investigated in this work. To address the limitation of only containing two severity levels in the NYC pedestrian–vehicle crash dataset, a supplementary dataset of highway crash data with three severity levels (no-injury, possible injury and evident injury) from 275 roadway segments in Washington State was used. Two MVPLN models were developed for the two datasets and the final models reported were developed by selecting the models that yield the lowest DIC values.

A new parallel sampling scheme was proposed and implemented in this study. A MATLAB code implementing the parallel sampling scheme was developed to estimate the MVPLN models in this paper. In the actual estimation, the parallelized scheme sped up the estimation process by 3 times when 8 parallel threads were used, and higher performance can be achieved if introducing more parallel threads in the computation. The goodness-of-fit measures such like DIC and log-likelihood evaluated at posterior mean of parameters were also computed. The improved estimation approach and tool developed in this paper will provide researchers and practitioners an efficient way to implementing MVPLN model in practice.

The MVPLN models offer theoretically sound approach to jointly model different severity levels' crashes and account for

the general structured correlation among different severity levels. We observed high correlation level for different severity levels in both of the two different crash datasets. For NYC pedestrian–vehicle crash dataset, the correlation value between pedestrian–vehicle fatal and severe injury crashes was 0.710. And for Washington State highway injury dataset, the correlation values among the three severity levels: no-injury, possible and evident injury are found to be 0.914 (no-injury and possible injury), 0.889 (no-injury and evident injury), 0.863 (possible and evident injury). Such high level of correlation from two different types of crash data clearly implied the existence of correlation among different severity levels. This demonstrates the necessity of jointly modeling crashes at multiple severity levels using multivariate count models. A comparison study of MVPLN model with univariate Poisson Lognormal, Poisson and Negative Binomial models was conducted, and superior fitting results were confirmed by investigating the DIC and log-likelihood values, which ruled out the applicability of using univariate models for both of the studied datasets. The superior fitting to the data will contribute to a better understanding of the underlying factors that impact the occurrence of crashes in both urban streets and highways.

Acknowledgments

We thank the CUBRC team, Kevin Majka and Dr. Alan Blatt, for the preparation of the NYC pedestrian-vehicle crash database. We would also like to thank Prof. Fred Mannering of Purdue University for providing us the Washington State highway injury dataset. The authors are solely responsible for the findings in this work.

References

Aguero-Valverde, J., 2013. Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: comparing the precision of crash frequency estimates. Accident Analysis and Prevention 50, 289–297.
Aguero-Valverde, J., Jovanis, P.P., 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. Transportation Research Record 2136, 82–91.
Aguero-Valverde, J., Jovanis, P.P., 2010. Spatial correlation in multilevel crash frequency models effects of different neighboring structures. Transportation Research Record 2165, 21–32.
Aziz, H. A., Ukkusuri, S. V., Hasan, S., 2013. Exploring the determinants of pedestrian-vehicle crash severity in New York City. Accident Analysis and Prevention 50, 1298-1309.
Barua, S., El-Basyouny, K., Islam, M.T., 2014. A full Bayesian multivariate count data model of collision severity with spatial correlation. Analytic Methods in Accident Research 3, 28–43.
Bijleveld, F.D., 2005. The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accident Analysis and Prevention 37 (4), 591–600.
Caliendo, C., De Guglielmo, M.L., Guida, M., 2013. A crash-prediction model for road tunnels. Accident Analysis and Prevention 55, 107–115.
Castro, M., Paleti, R., Bhat, C.R., 2012. A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. Transportation Research Part B 46 (1), 253–272.
Chib, S., Winkelmann, R., 2001. Markov chain Monte Carlo analysis of correlated count data. Journal of Business & Economic Statistics 19 (4), 428–435.
Chiou, Y.C., Fu, C., 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. Accident Analysis and Prevention 50, 73–82.
Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accident Analysis and Prevention 40 (4), 1257–1266.
El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention 41 (4), 820–828.
Eluru, N., Bagherib, M., Miranda-Morenoa, L., Fu, L., 2012. A latent class modeling approach for identifying vehicle driver injury severity factors at highway–railway crossings. Accident Analysis and Prevention 47, 119–127.
Hausman, J.A., Hall, B., Griliches, Z., 1984. Econometric models for count data with an application to the patents-R&D relationship. Econometrica 52 (4), 909–938.
Hensher, D., Greene, W., 2003. Mixed logit models: state of practice. Transportation 30 (2), 133–176.
Kim, K., Yamashita, E., 2002. Motor vehicle crashes and land use: empirical analysis from Hawaii. Transportation Research Record 1784, 73-79.
Kocherlakota, S., Kocherlakota, K., 1992. Bivariate Discrete Distributions. Marcel Dekker, New York.
LaScala, E.A., Gerber, D., Gruenewald, P.J., 2000. Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. Accident Analysis and Prevention 32 (5), 651–658.
Loukaitou-Sideris, A., Liggett, R., Sung, H.G., 2007. Death on the crosswalk: a study of pedestrian-automobile collisions in Los Angeles. Journal of Planning Education and Research 26 (3), 338–351.
Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention 40 (3), 964–975.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Analytic Methods in Accident Research 1, 1–22.
Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention 40 (1), 260–266.
Mohamed, M. G., Saunier, N., Miranda-Moreno, L. F., Ukkusuri, S. V., 2013. A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. Safety science 54, 27-37.
Narayanamoorthy, S., Paleti, R., Bhat, C.R., 2013. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. Transportation Research Part B 55, 245–264.
New York City Department of Transportation (NYCDOT), 2007. Safe Streets NYC: Traffic Safety Improvements in New York City. NYCDOT, New York June 2007.
Nikoloulopoulos, A.K., Karlis, D., 2010. Regression in a copula model for bivariate count data. J. Appl. Stat. 37 (9), 1555–1568.
Park, E.S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, 1–6.
Park, E.S., Park, J., Lomax, T.J., 2010. A fully Bayesian multivariate approach to before-after safety evaluation. Accident Analysis and Prevention 42 (4), 1118–1127.
Pei, X., Wong, S.C., Sze, N.N., 2011. A joint-probability approach to crash prediction models. Accident Analysis and Prevention 43 (3), 1160–1166.

Shin, H.S., Chen, G., Majka, K., Shorris, A., Ukkusuri, S., 2010. Pedestrian Fatality and Severe Injury Accidents in New York City. New York City Department of Transportation.
Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis 97 (1), 246–273.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B 64 (4), 583–639.
Ukkusuri, S.V., Hasan, S., Abdul Aziz, H.M., 2011. Random-parameter model to explain effects of built environment characteristics on pedestrian accident frequency. Transportation Research Record 2237, 98–106.
Wang, C., Quddus, M.A., Ison, S.G., 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. Accident Analysis and Prevention 43 (6), 1979–1990.
Wang, Y., Kockelman, K.M., 2013. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accident Analysis and Prevention 60, 71–84.
Winkelmann, R., 2000. Seemingly unrelated negative binomial regression. Oxford Bulletin of Economics and Statistics 62 (4), 553–560.
Winkelmann, R., 2008. Econometric Analysis of Count Data. Springer, Berlin Heidelberg.
Xiong, Y., Mannering, F.L., 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: a finite-mixture random-parameters approach. Transportation Research Part B 49, 39–54.
Yasmin, S., Eluru, N., Ukkusuri, S. V., 2014. Alternative ordered response frameworks for examining pedestrian injury severity in New York City. Journal of Transportation Safety and Security 6(4), 275-300.
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 Science 47 (3), 443–452.