Accident Analysis and Prevention 41 (2009) 820–828

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Collision prediction models using multivariate Poisson-lognormal regression

Karim El-Basyouny*, Tarek Sayed[1]

Dept. of Civil Engineering, University of British Columbia, 2002-6250 Applied Science Lane, Vancouver, BC, Canada V6T 1Z4

ARTICLE INFO

Article history:
Received 24 December 2008
Received in revised form 20 March 2009
Accepted 2 April 2009

Keywords:
Collision prediction models
Full Bayes estimation
Markov Chain Monte Carlo
Multivariate lognormal distribution
Multivariate identification of hot spots

ABSTRACT

This paper advocates the use of multivariate Poisson-lognormal (MVPLN) regression to develop models for collision count data. The MVPLN approach presents an opportunity to incorporate the correlations across collision severity levels and their influence on safety analyses. The paper introduces a new multivariate hazardous location identification technique, which generalizes the univariate posterior probability of excess that has been commonly proposed and applied in the literature. In addition, the paper presents an alternative approach for quantifying the effect of the multivariate structure on the precision of expected collision frequency. The MVPLN approach is compared with the independent (separate) univariate Poisson-lognormal (PLN) models with respect to model inference, goodness-of-fit, identification of hot spots and precision of expected collision frequency. The MVPLN is modeled using the WinBUGS platform which facilitates computation of posterior distributions as well as providing a goodness-of-fit measure for model comparisons. The results indicate that the estimates of the extra Poisson variation parameters were considerably smaller under MVPLN leading to higher precision. The improvement in precision is due mainly to the fact that MVPLN accounts for the correlation between the latent variables representing property damage only (*PDO*) and injuries plus fatalities (*I+F*). This correlation was estimated at 0.758, which is highly significant, suggesting that higher *PDO* rates are associated with higher *I+F* rates, as the collision likelihood for both types is likely to rise due to similar deficiencies in roadway design and/or other unobserved factors. In terms of goodness-of-fit, the MVPLN model provided a superior fit than the independent univariate models. The multivariate hazardous location identification results demonstrated that some hazardous locations could be overlooked if the analysis was restricted to the univariate models.

1. Introduction

The application of collision prediction models (CPMs) in assessing the safety of a road entity (intersection, road segment, etc.) has become a standard practice among safety researchers and practitioners (Hauer et al., 1988; Persaud and Dzbik, 1993; Lord, 2000; Miaou and Lord, 2003; Sawalha and Sayed, 2006a,b).

Data on the number of collisions at a particular site are usually available where the collisions are classified by severity (e.g., fatal, minor injury, major injury or property damage only), by the number of vehicles involved (e.g., single or multiple), and/or by the type of collision (e.g., angle, head-on, rear-end, sideswipe or pedestrian-involved), etc. In such cases, it is necessary to account for the likely correlations among collision counts at different levels of classification. These correlations may be caused by omitted variables, which can influence collision occurrence at all levels of classification, or from ignoring shared information in unobserved error terms. Several researchers such as Hydén (1987) and Sayed and Zein (1999) showed that severe conflicts and various collision severity levels did in fact share the same severity distribution. However, most studies have avoided the issue of correlations by using a frequency model to predict the total number of collisions and then conditional on collision occurring, severity can be predicted using techniques such as multinomial logit models, nested logit models, ordered probit models, sequential binary probit models, heteroskedastic multivariate generalized extreme value models and mixed logit models (e.g., see Yamamoto et al., 2008; Kim et al., 2008; Milton et al., 2008; and the references therein).

In contrast to the above two-stage approach of using a model to predict total collisions and a conditional model to predict severity, recent models for unconditional accident severity analysis have been proposed in the safety literature (Tunaru, 2002; Ma and Kockelman, 2006; Brijs et al., 2007; Park and Lord, 2007; Ma et al., 2008; Ye et al., 2009; Aguero-Valverde and Jovanis, 2009). These multivariate extensions are based on either multivariate Poisson (MVP) models (Tsionas, 2001; Karlis, 2003; Karlis and Meligkotsidou, 2005) or multivariate Poisson-lognormal (MVPLN) models (Chib and Winkelmann, 2001). The MVPLN regression is

* Corresponding author. Tel.: +1 604 716 4470.
E-mail addresses: basyouny@civil.ubc.ca (K. El-Basyouny), tsayed@civil.ubc.ca (T. Sayed).
[1] Tel.: +1 604 822 4379.

0001-4575/$ – see front matter.
doi:10.1016/j.aap.2009.04.005

preferred to the MVP approach for the analysis of multivariate collision count data because (i) it accounts for over-dispersion (extra Poisson variation), which is often observed in collision data; and (ii) it allows for a full general correlation structure. Since the classical estimation of the parameters of the MVP and MVPLN regression models is not straightforward, the Markov chain Monte Carlo (MCMC) simulation method (Gilks et al., 1996) is typically used in the applications.

This paper advocates the use of MVPLN for modeling collision count data. The MVPLN regression model presents an opportunity to corroborate the findings reported in the literature regarding the nature of the correlations across collision severity levels and their influence on safety analyses. The paper introduces a new multivariate hazardous location identification technique, which generalizes the univariate posterior probability of excess that has been commonly proposed and applied in the literature. In addition, the paper presents an alternative approach for quantifying the effect of the multivariate structure on the precision of expected collision frequency. The MVPLN approach is compared with the independent (separate) univariate PLN models with respect to model inference, goodness-of-fit, identification of hot spots and precision of expected collision frequency.

The development of the MVPLN model is undertaken using WinBUGS (Lunn et al., 2000) which is an open-source statistical software. WinBUGS is a flexible platform for the Bayesian analysis of complex statistical models using MCMC methods. The proposed implementation of the MVPLN regression model is relatively simpler than other MVPLN approaches using special codes written for the MATLAB and R computing environments (Park and Lord, 2007; Ma et al., 2008). WinBUGS has the additional advantage of providing the user with a goodness-of-fit measure (the deviance information criteria) that can be used for model comparisons.

2. Previous work

To distinguish between the collision prediction models that use multivariate explanatory variables to predict a univariate dependent variable (e.g., the total number of collisions) and those which involve multivariate dependent as well as independent variables, the former will be termed univariate CPMs, while the latter will be termed multivariate CPMs. It should be noted that both univariate and multivariate CPMs use multivariate independent variables (covariates). In view of this terminology, the popular univariate approach for developing CPMs uses the Poisson-gamma hierarchy, which leads to the negative binomial regression model (e.g., Poch and Mannering, 1996; Hauer, 1997; Hinde and Demetrio, 1998; Lord, 2000; Miaou and Lord, 2003). The PLN regression represents a viable alternative for modeling the extra-Poisson variation (Kim et al., 2002; Miranda-Moreno, 2006).

The majority of CPMs were developed using models with a fixed dispersion parameter. Such an assumption was challenged and various dispersion parameter relationships were examined (Heydecker and Wu, 2001; Miaou and Lord, 2003; Miranda-Moreno et al., 2005; El-Basyouny and Sayed, 2006; Mitra and Washington, 2007; Lord and Park, 2008).

There is extensive research addressing the problem of observing excessive zeroes in collision data (e.g., Shankar et al., 2004; Lee and Mannering, 2002; Kumara and Chin, 2006; Qin et al., 2004; Warton, 2005; Lord et al., 2005, 2007). Another problem of considerable interest is the development of CPMs using data characterized by a low sample mean, especially if combined with a small sample size, (e.g., Maycock and Hall, 1984; Fridstrøm et al., 1995; Wood, 2002; Lord, 2006). Other modeling techniques have also been proposed in the literature advocating the use of random parameter negative binomial regression model (Anastasopoulos and Mannering, 2009) or a two-state Markov Switching model (Malyshkina et al., 2008) to analyze collision frequencies.

Multivariate extensions of the simple Poisson distribution to account for the correlations among different count processes (such as counts of different crash severities) have been proposed in the literature. For instance, a model that assumes the same nonnegative covariance term for all pairs has been considered by Tsionas (2001) and Karlis (2003). Ma and Kockelman (2006) adapted a multivariate Poisson (MVP) regression approach to assess the effects of different covariates on collision counts at different severity levels. The assumption of equal covariance terms has been relaxed by Karlis and Meligkotsidou (2005), but the restriction to nonnegative covariance terms was still maintained. Brijs et al. (2007) used a similar non-regression model for ranking hazardous sites in Belgium.

Multivariate zero-inflated Poisson (MVZIP) models were also considered (Li et al., 1999). Since the MVP regression models do not allow for over-dispersion, a multivariate negative binomial (MVNB) model was used to investigate the simultaneity of fatality and injury collision outcomes (Ladron de Guevara and Washington, 2004). A major drawback of the MVNB models is its inability to tolerate negative correlations.

To address the shortcomings of MVP and MVNB, Chib and Winkelmann (2001) developed the MVPLN regression approach. The MVPLN model does not only account for over-dispersion, but also it has a fairly general correlation structure allowing for (i) different covariance terms; and (ii) the possibility of negative correlations, which cannot be ruled out unquestionably for all classifications.

Recently, applications of the MVPLN to model collision count data at different levels of severity were undertaken. Tunaru (2002) used an MVPLN non-regression model implemented via WinBUGS for analyzing multiple response count data (classified by severity and the number of vehicles involved) taking into account complex correlation structures.

Park and Lord (2007) used MATLAB codes MATLAB (2006), tailored according to the MCMC algorithm of Chib and Winkelmann (2001), for jointly modeling collision frequency by severity using data from 451 three-leg un-signalized intersections in California obtained through the Highway Safety Information System (HSIS). Their results showed promise toward obtaining more accurate estimates by accounting for correlations in the multivariate collision counts and over-dispersion.

Ma et al. (2008) coded a Gibbs sampler and two Metropolis–Hastings algorithms (Gilks et al., 1996) in the R language for the prediction of collision counts by severity using data collected from Washington State through the HSIS on 7773 rural two-lane roadways in the Puget Sound region. Their results indicated that the MVPLN approach offered better predictions than those from univariate Poisson and negative binomial models. The results indicated that there were statistically significant correlations between (as well as over-dispersion in) collision counts at different levels of injury severity. Several recommendations for highway safety treatments and design policies were suggested, e.g., they concluded that wide lanes and shoulders are key for reducing collision frequencies, as are longer vertical curves.

Ye et al. (2009) developed a simultaneous equations model of collision frequencies by collision type for a sample of 165 rural intersections in 38 counties of Georgia. The model was formulated using multivariate Poisson regression with multivariate normal heterogeneity. A simulation-based maximum likelihood approach was used to estimate the parameters. The model estimation results support the notion of the presence of significant common unobserved factors across collision types, although the impact of these factors on parameter estimates was found to be rather modest.

Aguero-Valverde and Jovanis (2009) used Full Bayes MVPLN models implemented via WinBUGS to estimate the expected col-

lision frequency for different collision severity levels using data on 6353 rural two-lane segments in central Pennsylvania. They found high correlations among collision severities, with highest values between contiguous severity levels. In terms of goodness-of-fit, the multivariate model outperformed the independent univariate PLN models and provided improved precision for collision frequency estimates. The multivariate estimates were combined with cost data from PennDOT to rank sites for safety improvements via expected collision cost and excess expected cost per segment. The results showed that the multivariate-based top ranked segments have consistently higher costs and excess costs than the univariate estimates as a result of higher multivariate estimates of fatalities and major injuries. These higher estimates resulted in different rankings for the multivariate and independent models.

3. Methodology

3.1. The multivariate Poisson-lognormal model

For a set of data on road collisions at n locations, where the collisions at each location are classified into K categories, define the vector $y_i = (y_{i1}\ y_{i2} \ldots y_{iK})'$, where y_{ik} denote the number of collisions at the ith location in category k. It is assumed that the y_i are independently distributed and that the Poisson distribution of y_{ik}, given λ_{ik}, is

$$f(y_{ik}|\lambda_{ik}) = \lambda_{ik}^{y_{ik}} \frac{e^{-y_{ik}}}{y_{ik}!}, \quad i = 1, 2, \ldots, n, \quad k = 1, 2, \ldots, K. \tag{1}$$

To model extra variation, assume further that $\ln(\lambda_{ik}) = \ln(\mu_{ik}) + \varepsilon_{ik}$, where

$$\ln(\mu_{ik}) = \beta_{k0} + \beta_{k1}X_{i1} + \cdots + \beta_{kJ}X_{ij}, \tag{2}$$

the X_{ij} denote relevant traffic and road characteristics and the ε_{ik} denote multivariate normal errors distributed as $\varepsilon_i \sim N_k(0, \Sigma)$, where

$$\varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \cdots \\ \varepsilon_{iK} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1K} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2K} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{K1} & \sigma_{K2} & \ldots & \sigma_{KK} \end{pmatrix}.$$

Let X and β_k denote the matrix of covariates and the vector of regression coefficients, respectively, and let β denote the set $\{\beta_1, \beta_2, \ldots, \beta_K\}$. Thus, given (X, β, Σ), the λ_i are independently distributed as

$$f(\lambda_i|X, \beta, \Sigma) = \frac{\exp\{-0.5(\lambda_i^* - \mu_i^*)'\Sigma^{-1}(\lambda_i^* - \mu_i^*)\}}{(2\pi)^{K/2}\left(\prod_{k=1}^{K}\lambda_{ik}\right)|\Sigma|^{1/2}}, \tag{3}$$

which is a K-dimensional log-normal distribution, where

$$\lambda_i = (\lambda_{i1} \quad \lambda_{i2} \cdots \lambda_{iK})', \quad \lambda_i^* = (\ln(\lambda_{i1}) \quad \ln(\lambda_{i2}) \cdots \ln(\lambda_{iK}))',$$
$$\mu_i = (\mu_{i1} \quad \mu_{i2} \cdots \mu_{iK})', \quad \mu_i^* = (\ln(\mu_{i1}) \quad \ln(\mu_{i2}) \cdots \ln(\mu_{iK}))'.$$

It should be noted that the univariate PLN models are obtained as the special case where Σ is a diagonal matrix.

3.2. Prior distributions

Let λ denote the set $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. To obtain the full Bayes estimates of (λ, β, Σ), it is required to specify prior distributions for the hyper parameters (β, Σ). Prior distributions are meant to reflect to some extent prior knowledge about the parameters of interest. If such prior information is available, it should be used to formulate the so-called informative priors. The specification of informative priors for generalized linear models was dealt with by Bedrick et al. (1996), who considered conditional means priors as well as data augmentation priors of the same form as the likelihood and showed that such priors result in tractable posteriors. A good discussion on the elicitation of priors in collision data analysis can be found in Schluter et al. (1997).

In the absence of sufficient prior knowledge of the distributions for individual parameters, uninformative (vague) proper prior distributions are usually specified. The most commonly used priors are diffused normal distributions (with zero mean and large variance) for the regression parameters and a Wishart (P, r) prior for Σ^{-1}, where P and $r \geq K$ represent the prior guess at the order of magnitude of the precision matrix Σ^{-1} and the degrees of freedom, respectively. The parameterization of the Wishart probability density function is

$$f(\Sigma^{-1}|P, r) = |P|^{K/2}|\Sigma^{-1}|^{(r-K-1)/2}\exp\{-0.5Tr(P\Sigma^{-1})\}.$$

Choosing $r = K$ as the degrees of freedom corresponds to vague prior knowledge (Spiegelhalter et al., 1996; Tunaru, 2002).

3.3. Posterior distributions

Let $f(\beta) \equiv \prod_{k,j} N(0,10{,}000)$ and $f(\Sigma^{-1}) \equiv$ Wishart (P, K) denote the hyper-prior distributions for the regression parameters and the precision matrix, respectively, Further, let y denote the set $\{y_1, y_2, \ldots, y_n\}$. Given X, the joint distribution of $(y, \lambda, \beta, \Sigma)$ is

$$f(y, \lambda, \beta, \Sigma|X) = \frac{f(\beta)f(\Sigma^{-1})}{(2\pi)^{nK/2}|\Sigma|^{n/2}\left(\prod_{i=1}^{n}\prod_{k=1}^{K}y_{ik}!\right)}g(y, \lambda, X, \beta, \Sigma),$$

where

$$g(y, \lambda, X, \beta, \Sigma) = \exp\left\{-0.5\sum_{i=1}^{n}\left((\lambda_i^* - \mu_i^*)'\Sigma^{-1}(\lambda_i^* - \mu_i^*) + \sum_{k=1}^{K}[(y_{ik} - 1)\ln(\lambda_{ik}) - \lambda_{ik}]\right)\right\}.$$

The posterior distributions are given by

$$f(\lambda_i|y, X, \beta, \Sigma) = \frac{g_i(y_i, \lambda_i, X, \beta, \Sigma)}{\int_0^\infty\int_0^\infty\cdots\int_0^\infty g_i(y_i, \lambda_i, X, \beta, \Sigma)d\lambda_{i1}d\lambda_{i2}\ldots d\lambda_{iK}}, \tag{4}$$

where

$$g_i(y_i, \lambda_i, X, \beta, \Sigma) = \exp\left\{-0.5\lambda_i^* - \mu_i^*)'\Sigma^{-1}(\lambda_i^* - \mu_i^*) + \sum\nolimits_{k=1}^{K}[(y_{ik} - 1)\ln(\lambda_{ik}) - \lambda_{ik}]\right\},$$

$$f(\beta|y, \lambda, X, \Sigma) = \frac{g(y, \lambda, X, \beta, \Sigma)f(\beta)}{\int_{-\infty}^\infty\int_{-\infty}^\infty\cdots\int_{-\infty}^\infty g(y, \lambda, X, \beta, \Sigma)f(\beta)d\beta_{10}d\beta_{11}\ldots d\beta_{KJ}}, \tag{5}$$

and

$$f(\Sigma^{-1}|y, \lambda, X, \beta) \equiv Wishart\left(P + \sum_{i=1}^{n}(\lambda_i^* - \mu_i^*)(\lambda_i^* - \mu_i^*)', K + n\right). \tag{6}$$

3.4. Full Bayes estimation

The posterior distributions needed in the full Bayes approach can be obtained using MCMC sampling. The Wishart distribution in (6) can be sampled using a Gibbs sampler. In contrast, the posterior distributions in (4) and (5) are not standard and require the

Metropolis–Hastings algorithm (Park and Lord, 2007; Ma et al., 2008).

In this paper, the posterior distributions are sampled using the MCMC techniques available in WinBUGS 2.2.0; the windows interface of OpenBUGS. The techniques generate sequences (chains) of random points, whose distributions converge to the target posterior distributions. A sub-sample is used to monitor convergence and then excluded as a burn-in sample. The remaining iterations are used for parameter estimation, performance evaluation and inference.

Monitoring convergence is important because it ensures that the posterior distribution has been "found". Thereby indicating when parameters sampling should begin. To check convergence, two or more parallel chains with diverse starting values are tracked to ensure full coverage of the sample space. Convergence of multiple chains is assessed using the Brooks–Gelman–Rubin (BGR) statistic (Brooks and Gelman, 1998). A value under 1.2 of the BGR statistic indicates convergence. Convergence is also assessed by visual inspection of the MCMC trace plots for the model parameters as well as by monitoring the ratios of the Monte Carlo errors relative to the respective standard deviations of the estimates; as a rule of thumb these ratios should be less than 0.05.

3.5. Model comparisons

The Deviance Information Criteria (DIC) is used for model comparisons (Spiegelhalter et al., 2002). As a goodness of fit measure, DIC is a Bayesian generalization of Akaike's Information Criteria (AIC) that penalizes larger parameter models. According to Spiegelhalter et al. (2005), it is difficult to determine what would constitute an important difference in DIC. Very roughly, differences of more than 10 might definitely rule out the model with the higher DIC. Differences between 5 and 10 are considered substantial. If the difference in DIC is less than 5, and the models make very different inferences, then it could be misleading just to report the model with the lowest DIC.

To show that DIC is additive under independent models and priors, let $f(y|\theta)$ and $f(y)$ denote the conditional and marginal distributions of y, where θ denote the vector of parameters associated with y. Then, $DIC = \bar{D} + p$, where $p = \bar{D} - D(\bar{\theta})$, $\bar{\theta} = E[\theta|y]$ and $\bar{D} = E[D(\theta)|y]$ are the posterior means of θ and the Bayesian deviance

$$D(\theta) = -2\ \ln\{f(y|\theta)\} + 2\ln\{f(y)\}. \tag{7}$$

For K collision categories, let y and θ be partitioned as $(y_1,\ldots, y_k)$ and $(\theta_1,\ldots, \theta_k)$. Define $DIC_k = \bar{D}_k + p_k$, $p_k = \bar{D}_k - D_k(\bar{\theta}_k)$, $\bar{D}_k = E[D_k(\theta_k)|y_k]$, $\bar{\theta}_k = E[\theta_k|y_k]$ and $D_k(\theta_k) = -2\ln\{f(y_k|\theta_k)\} + 2\ln\{f(y_k)\}$. Under independent models and priors, we have that $f(y|\theta) = \prod_{k=1}^{K} f(y_k|\theta_k)$ and $f(y) = \prod_{k=1}^{K} f(y_k)$. These multiplicative conditional and marginal distributions of y translate additively in the Bayesian deviance (7) leading to $DIC = \sum_{k=1}^{K} DIC_k$.

4. Multivariate identification

The treatment in this section is not meant to account for the extensive literature on the selection of hazardous locations, but rather to illustrate the generalization to the multivariate setting of some of the techniques that were developed for univariate analysis.

Several hazardous location identification techniques exist in the literature. These include the probability of worst site, posterior distribution of the ranks of the collisions costs or expected number of collisions, or through expected collision cost and excess expected cost (Tunaru, 2002; Brijs et al., 2007; Aguero-Valverde and Jovanis, 2009).

However, assigning costs to different types of injury can be controversial for a variety of reasons (Brijs et al., 2007). These reasons include ethical arguments (e.g., can we assign a cost to a human life?) and economic arguments (what are the quantities that must be measured to estimate the cost of a seriously injured person?). To avoid such difficulties, an alternative approach of selecting hot spots is adopted which is based on the posterior probability of excess (Higle and Witkowski, 1988; Sayed and Abdelwahab, 1997; Heydecker and Wu, 2001). The Bayesian posterior probability that a site has an excessive mean collision frequency provides an indication of sites at which the future collision record is expected to be worse than usual. In the univariate analysis, this probability is expressed as:

$$\int_{\mu 0}^{\infty} f(\lambda_i|y, X, \beta, \sigma^2)d\lambda_i > 1-\delta, \quad i = 1, 2, \ldots, n, \tag{8}$$

where σ^2, μ_0 and δ denote the extra-Poisson variation, an upper limit of the "acceptable" mean number of collisions and a threshold value between 0 and 1, respectively. In practice, μ_0 is typically specified as either the median or mean of the prior distribution, whereas δ has been arbitrarily assumed to equal 0.05 (Higle and Witkowski, 1988; Sayed and Abdelwahab, 1997).

Under the multivariate Poisson-lognormal model, the criterion (8) generalizes to

$$\int_{\mu_{10}}^{\infty}\int_{\mu_{20}}^{\infty}\cdots\int_{\mu K0}^{\infty} f(\lambda_i|y, X, \beta, \Sigma)d\lambda_{i1}d\lambda_{i2}\ldots d\lambda_{iK} > 1-\delta,$$
$$i = 1, 2, \ldots, n, \tag{9}$$

where $f(\lambda_i|y, X, \beta, \Sigma)$ is given by (4) and μ_{k0} denotes an upper limit of the "acceptable" mean number of collisions in category k. It should be noted that the univariate Eq. (8) is the special case $K = 1$. The evaluation of the multiple integral in (9) is rather complicated. Thereby, the following approximation, due to Clayton and Kaldor (1987), is proposed to simplify the computation of (9),

$$f(\lambda_i^*|y, X, \beta, \Sigma) \approx N_K(mi, Si) \tag{10}$$

where

$$m_i = S_i\left(\Sigma^{-1}\mu_i^* + \begin{pmatrix} (y_{i1}+0.5)\ln(y_{i1}+0.5) - 0.5 \\ (y_{i2}+0.5)\ \ln(y_{i2}+0.5) - 0.5 \\ \cdots \\ (y_{iK}+0.5)\ \ln(y_{iK}+0.5) - 0.5 \end{pmatrix}\right), \tag{11}$$

and

$$S_i = \left(\Sigma^{-1} + \begin{pmatrix} y_{i1}+0.5 & 0 & \ldots & 0 \\ 0 & y_{i2}+0.5 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & y_{iK}+0.5 \end{pmatrix}\right)^{-1}. \tag{12}$$

The approximation (10) can be used in the Bayesian probability of excess criterion (9) to compare the univariate and multivariate approaches of selecting hot spots.

For univariate PLN models ($K = 1$) let μ_0^* denote the natural logarithm of the average of the prior means. Then, the ith intersection would be selected as a hazardous location whenever

$$\Phi(\mu_{i0}^*) < \delta, \tag{13}$$

where Φ denotes the univariate standard normal distribution function and

$$\mu_{i0}^* = \frac{\mu_0^* - m_i}{\sqrt{S_i}}.$$

The standardized multivariate normal distribution function can be used in a similar procedure to select hot spots under MVPLN. Thus, using a similar notation, the ith intersection would be selected as a hazardous location whenever

$$\Phi_K(\mu_{i10}^*, \mu_{i20}^*, \ldots, \mu_{iK0}^*, \hat{R}) < \delta, \tag{14}$$

where Φ_k denotes the multivariate standard normal distribution function,

$$\mu^*_{ik0} = \frac{\mu^*_{k0} - m_{ik}}{\sqrt{S_{ikk}}}, \quad k = 1, 2, \ldots, K,$$

and $\hat{R}$ denotes the correlation matrix corresponding to $\hat{\Sigma}$, the Bayesian estimate of the covariance matrix Σ.

5. Data and model description

A total of $n = 99$ signalized intersections in the city of Edmonton were investigated for the purpose of developing collision prediction models relating the safety of urban 4-leg intersections to their traffic flows. This data set was used as a reference group for the evaluation of the City of Edmonton's Intersection Safety Camera program (Sayed and de Leur, 2007).

The data on collision frequencies and traffic volumes within the intersections were obtained from the city of Edmonton. The traffic and collision data were provided for a 3-year period before the implementation of the program and were selected to ensure that no changes in either the intersection layout or traffic volumes occurred over the course of the 3 years. The term "collision" as defined by Edmonton's Transportation Department, refers to reportable on-street collisions that do not occur on private property, include at least one motor vehicle, and result in injury, at least $1,000 in property damage, or both. The database does not include non-vehicular collisions (e.g., cyclist hitting a pedestrian) and includes collisions that are forwarded to the department by the police service within a specified period. The term "intersection collisions" include collisions that occur inside and 10 m past the legally defined limits of the outer crosswalk lines of intersecting roads and include right-turn cut-offs. The collisions were classified into $K = 2$ categories: property damage only (*PDO*), $k = 1$, and injuries and fatalities ($I + F$), $k = 2$.

Average values for the traffic volumes (over the 3 years period) are used to build CPMs for predicting aggregate number of collisions

$$\ln(\mu_{ik}) = \beta_{k0} + \beta_{k1} X_{i1} + \beta_{k2} X_{i2}, \tag{15}$$

where $X_{i1} = \ln(V_{i1})$, V_{i1} = AADT at the major approach, and $X_{i2} = \ln(V_{i2})$, V_{i2} = AADT at the minor approach. The aggregation is justified on several grounds. For instance, an aggregate CPM was found to perform well compared with CPMs developed to handle temporal correlation (Lord and Persaud, 2000). Also, the aggregation of collisions over a period of reasonable length helps to avoid confounding effects and such a phenomenon like regression-to-the-mean (Cheng and Washington, 2005). A statistical summary of the data is shown in Table 1.

To explore the potential relationships between *PDO*/$I + F$ and major and minor traffic volumes, Fig. 1 displays the three-dimensional relationships via wireframe plots.

Fig. 1 lends support to the AADT functional form (15), as the relationship between each of *PDO* and $I + F$ and major and minor traffic volumes, is effectively linear on the logarithmic scale, except for a very few cases where there were no collisions. The AADT functional form (15) is popular among practitioners and is often preferred over models that include several covariates because it can be easily re-calibrated (Lord et al., 2008). Further, it will be adopted for estimating safety performance in the forthcoming highway safety manual (Hughes et al., 2005). It should be noted that AADT models may suffer from omitted variables bias since many non-flow related factors are known to affect collision frequency. Nevertheless, such models are well recognized in the traffic safety literature.

6. Results and discussion

6.1. Model results

Tables 2 and 3 summarize the parameter estimates and their associated statistics using the univariate and multivariate Poisson-lognormal models. In contrast to the univariate PLN which has multivariate explanatory variables, the MVPLN involves multivariate dependent as well as independent variables. It should also be noted that while MVPLN can handle more than two collision categories, the current application involves only two severity levels leading to a bivariate PLN.

The posterior summaries in Tables 2 and 3 were obtained via two chains with 20,000 iterations 10,000 of which were excluded as a burn-in sample using WinBUGS. A Wishart prior with an identity scale matrix and two degrees of freedom was adopted (Chib and Winkelmann, 2001; Congdon, 2006). Examination of the BGR statistics, ratios of the Monte Carlo errors relative to the standard deviations of the estimates and trace plots for all model parameters indicated convergence.

The results of Tables 2 and 3 show that: (i) the parameter estimates are significant, as the 95% credible intervals are bounded away from zero; (ii) the regression coefficients are positive, indicating an increase in the mean collision frequency with traffic volumes; (iii) the impact of traffic volumes at the major approach is larger than that at the minor approach; and (iv) traffic volumes have a larger effect on *PDO* collisions than on $I + F$ collisions.

Also, Tables 2 and 3 reveal small differences in the estimates of the regression parameters between the univariate and multivariate models. However, the estimates of the extra Poisson variation parameters (σ_{11} and σ_{22}) were considerably smaller under MVPLN. In fact, the MVPLN 95% credible intervals were entirely shifted to the left of those obtained under the univariate PLN models, for both *PDO* and $I + F$.

For $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, K$, the mean and variance of Y_{ik} are given by

$$E(Y_{ik}) = \mu_{ik} \exp(0.5\sigma_{kk}), \tag{16}$$

and

$$Var(Y_{ik}) = E(Y_{ik}) + [E(Y_{ik})]^2(\exp(\sigma_{kk}) - 1). \tag{17}$$

Since the second term in Eq. (17) dominates the first, we have that

$$Var(Y_{ik}) \cong [E(Y_{ik})]^2(\exp(\sigma_{kk}) - 1),$$

leading to the ratio

$$\frac{Var(Y_{ik}|MVPLN)}{Var(Y_{ik}|PLN)} \cong \frac{(\exp(\hat{\sigma}_{kk}) - 1)}{(\exp(\tilde{\sigma}_{kk}) - 1)}, \tag{18}$$

where $\hat{\sigma}_{kk}$ and $\tilde{\sigma}_{kk}$ are the posterior estimates of σ_{kk} under MVPLN and PLN, respectively. As precision is inversely proportional to the variance of expected collision frequency, it is estimated that MVPLN is more than twice as precise as the independent univariate PLN models; the actual precision ratios were computed via Eq. (18) as 2.5 and 2.1 for *PDO* and $I + F$, respectively. The precision of the expected collision frequencies can be estimated also using the standard deviations via Eq. (17). Fig. 2 displays these standard deviations by severity and model.

Fig. 2 reveals that (i) the standard deviations of $I + F$ are smaller than those of *PDO*, (ii) the standard deviations of the multivariate model are uniformly smaller than those of the PLN models, (iii) the relationship between the two sets of standard deviations appears to be linear for the two severity levels, and (iv) the standard deviations of the multivariate model were decreased 41% and 36% for *PDO* and $I + F$, respectively. With a larger data set for rural two-lane segments in central Pennsylvania, Aguero-Valverde and

Table 1
Statistical summary of data set (99 intersections).

	Minimum	Maximum	Mean	Standard deviation
AADT on major approach	5,642	61,744	28,010	9,705
AADT on minor approach	2,809	29,128	16,043	6,564
Number of property damage only (*PDO*) collisions	0	126	43.13	24.96
Number of injuries and fatalities (*I* + *F*) collisions	0	78	26.99	15.59

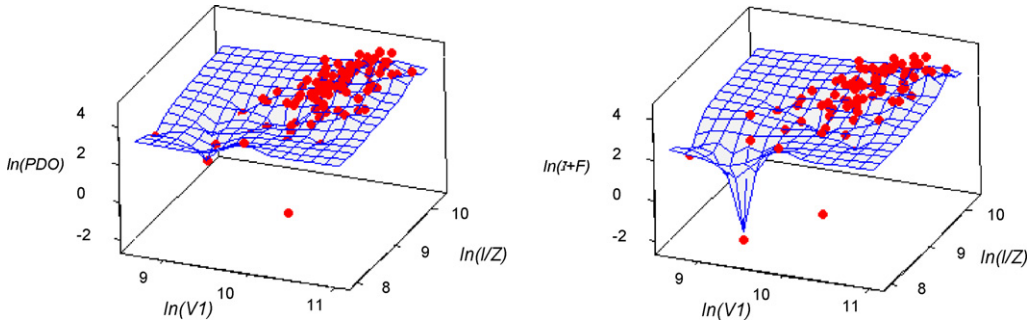

Fig. 1. Left: wireframe plot of ln(*PDO*) vs. ln(V_1) and ln(V_2). Right: wireframe plot of ln(*I* + *F*) vs. ln(V_1) and ln(V_2). *PDO*: property damage only, *I* + *F*: injuries and fatalities.

Table 2
Univariate models' statistics.

Univariate Poisson-lognormal	Estimate	Standard deviation	95% credible intervals	
			Lower limit	Upper limit
Property damage only (*PDO*)				
Intercept	−10.590	1.282	−13.090	−7.950
Ln(V_1) (major AADT)	0.927	0.131	0.667	1.188
Ln(V_2) (minor AADT)	0.496	0.097	0.307	0.686
σ_{11}	0.361	0.036	0.297	0.437
DIC	730.3			
Injuries and fatalities (*I* + *F*)				
Intercept	−9.267	1.447	−12.180	−6.440
Ln(V_1) (major AADT)	0.754	0.148	0.480	1.042
Ln(V_2) (minor AADT)	0.493	0.105	0.292	0.706
σ_{22}	0.415	0.043	0.339	0.507
DIC	684.3			

Jovanis (2009) obtained reductions in the standard deviations of 20.5% (*PDO*), 10.5% (minor injuries), 21.5% (moderate injuries), 48.1% (major injuries) and 40.7% (fatalities).

The improvement in precision is due mainly to the fact that MVPLN accounts for the correlation between the latent variables λ_{i1} (*PDO*) and λ_{i2} (*I* + *F*). This correlation (ρ) was estimated at 0.758, which is highly significant. In essence, higher *PDO* rates are associated with higher *I* + *F* rates, as the collision likelihood for both types is likely to rise due to the same deficiencies in roadway design and/or other unobserved factors.

As noted in previously, AADT-only models may suffer from omitted variables bias since the unobserved heterogeneity from other factors known to influence collision frequency (such as number of lanes, signal-control timing, speed limits, etc.) ends up in the

Table 3
Multivariate models' statistics.

Multivariate Poisson-lognormal	Estimate	Standard deviation	95% credible intervals	
			Lower limit	Upper limit
Property damage only (*PDO*)				
Intercept	−10.680	1.347	−13.420	−8.250
Ln(V_1) (major AADT)	0.902	0.138	0.643	1.179
Ln(V_2) (minor AADT)	0.531	0.104	0.320	0.723
σ_{11}	0.163	0.030	0.113	0.230
Injuries and fatalities (*I* + *F*)				
Intercept	−9.346	1.565	−12.410	−6.240
Ln(V_1) (major AADT)	0.742	0.159	0.435	1.059
Ln(V_2) (minor AADT)	0.513	0.121	0.276	0.761
σ_{22}	0.217	0.042	0.145	0.312
Covariance				
σ_{12}	0.143	0.029	0.092	0.208
Correlation				
$\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$	0.758	0.056	0.634	0.852
DIC	1373			

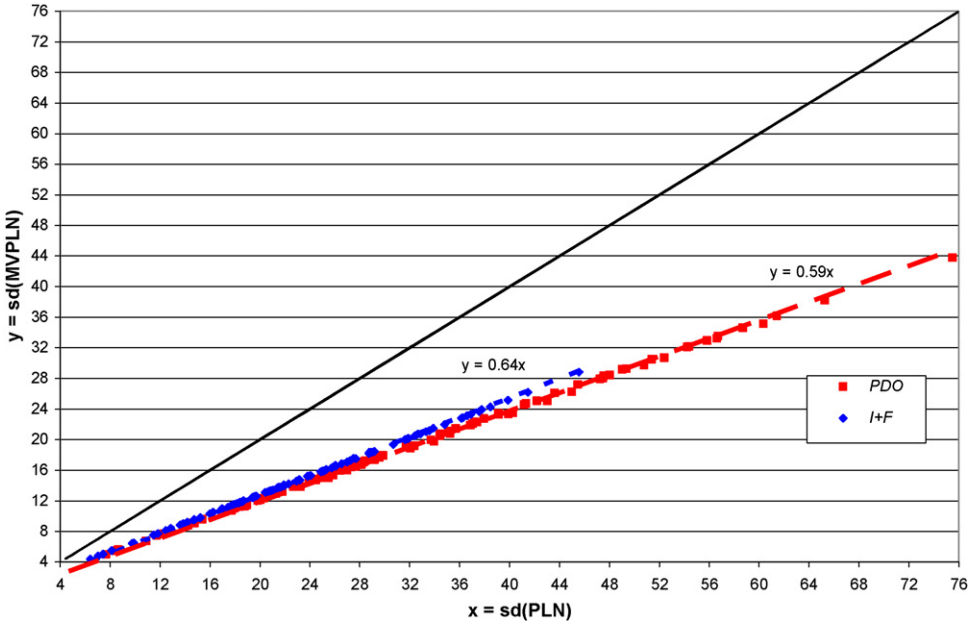

Fig. 2. Standard deviations (S.D.) of expected collision frequencies by severity and model. *PDO*: property damage only, *I + F*: injuries and fatalities.

correlation structure, thus affecting the estimated correlation. To study the magnitude of this effect on the correlation, the literature was consulted where several independent variables were used to develop multivariate CPMs.

Park and Lord (2007) reported a number of correlations between *PDO* and four severity levels. Upon combining these 4 severity levels in one category (*I + F*) and using the covariance matrix, the correlation between *PDO* and *I + F* was estimated at 0.8608. A similar calculation using the results of Aguero-Valverde and Jovanis (2009) produced a correlation estimate of 0.58. Ma et al. (2008) did not report the variances associated with the various severity levels. However, assuming comparable precisions, their results produced a rough correlation estimate of 0.44.

These results indicate that the current estimate of 0.758 is in the upper range of the estimates reported in the literature. Although the inclusion of additional relevant covariates is expected to reduce the correlation, it is conjectured to remain significant with crucial inferential impact.

The DIC statistics were 730.3, 684.3 and 1373 under the *PDO* PLN, *I + F* PLN and MVPLN, respectively. Thus, the MVPLN model provides a superior fit over the two univariate models as its (multivariate) DIC is much less than the sum of their (univariate) DICs; a very significant drop off of 41.6.

6.2. Multivariate identification of hazardous locations

Hazardous locations were identified via Eqs. (13) and (14) using $\delta = 0.10$, 0.05 and 0.01. For $K = 2$, the standardized bivariate normal distribution function PROBBNRM (SAS, 2002) was used in the implementation of Eq. (14). The PROBBNRM function required three input arguments $(\mu^*_{10} - m_{i1})/\sqrt{S_{i11}}$, $(\mu^*_{20} - m_{i2})/\sqrt{S_{i22}}$, and $\hat{\rho} = 0.758$. The outcomes appear in Table 4, where the results of the univariate Total PLN model were obtained through direct simulation of the sum *PDO + I + F*.

According to Table 4, the numbers of intersections identified as hot spots were 31, 29, 35 and 41 under *PDO* PLN, *I + F* PLN, Total PLN and MVPLN, respectively, for $\delta = 0.10$. The corresponding numbers were 30, 28, 32 and 37, for $\delta = 0.05$, and 22, 20, 28 and 30 for $\delta = 0.01$. As expected the numbers of hot spots was reduced as the threshold value decreased. However, several locations identified as hazardous by the MVPLN model were missed by the univariate models regardless of the choice of threshold value. It should also be noted that none of the intersections that were not identified as hazardous by MVPLN were identified by any of the univariate models except for one location under the *PDO* PLN model for $\delta = 0.10$ and 0.05.

The results of Table 4 demonstrate that some hazardous locations can be overlooked if the analysis was restricted to the univariate models. These results are consistent with those of the cost ranking approach of Aguero-Valverde and Jovanis (2009) in that the MVPLN and independent PLN models yield different identification/ranking of hot spots.

The advantage of MVPLN over separate univariate analyses is illustrated in Fig. 3, where posterior information is also contrasted to prior knowledge. The figure displays two credible ellipses and a credible rectangle for $(\ln(\lambda_{PDO}),\ln(\lambda_{I+F}))$, with 95% confidence. Fig. 3 was prepared using the data for Site 56, which was identified as hazardous by MVPLN and *I + F* PLN but neither by *PDO* PLN nor by Total PLN. Similar patterns were observed for other identified sites.

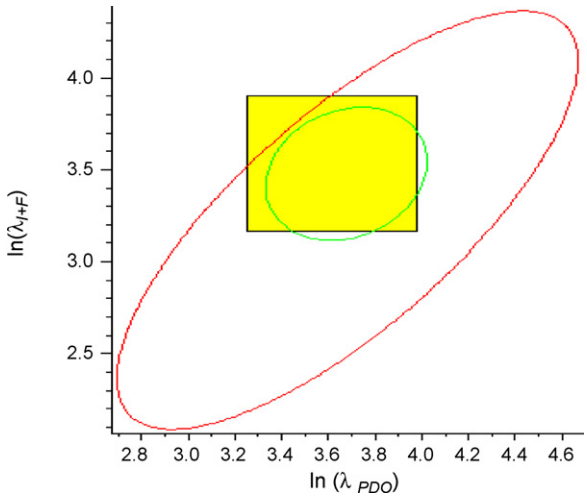

Fig. 3. 95% credible ellipses and rectangle for $(\ln(\lambda_{PDO}),\ln(\lambda_{I+F}))$, using the data for Site 56. *PDO*: property damage only, *I + F*: injuries and fatalities.

Table 4
Number of hot spots selected under different PLN models.

	Hazardous?	Multivariate PLN					
		$\delta = 0.10$		$\delta = 0.05$		$\delta = 0.01$	
		No	Yes	No	Yes	No	Yes
Univariate PLN							
Property damage only (*PDO*)	No	57	11	61	8	69	8
	Yes	1	30	1	29	0	22
Injuries and fatalities (*I* + *F*)	No	58	12	62	9	69	10
	Yes	0	29	0	28	0	20
Total	No	58	6	62	5	69	2
	Yes	0	35	0	32	0	28

The larger ellipse corresponds to the prior distribution in Eq. (3). The area inside that ellipse corresponds to 95% probable values for $(\ln(\lambda_{PDO}),\ln(\lambda_{I+F}))$ before the sample data were collected. The smaller ellipse corresponds to the posterior distribution in Eq. (10). The area inside that ellipse corresponds to 95% probable values for $(\ln(\lambda_{PDO}),\ln(\lambda_{I+F}))$ after incorporating the sample data into the analysis. It is clear that the sample information has enhanced prior knowledge and resulted in a more focused "posterior" ellipse that is a (small) subset of the "prior" ellipse.

The rectangle was constructed via the Bonferroni method using the two credible intervals obtained from the separate univariate models, *PDO* PLN and *I* + *F* PLN. The shaded area inside that rectangle corresponds to 95% probable values for $(\ln(\lambda_{PDO}),\ln(\lambda_{I+F}))$ obtained by combining the two separate univariate posterior intervals. The posterior ellipse is nearly contained in the rectangle indicating that some values that are improbable under the MVPLN posterior analysis would be probable under the separate univariate analyses and vice versa. Also, some of the probable values under the rectangle seem to contradict prior knowledge as they lie outside the prior ellipse.

7. Conclusions and future research

A MVPLN regression was used to jointly analyze a sample of collision counts classified by two severity levels (*PDO* and *I* + *F*) at 99 urban intersections. To illustrate the importance of the multivariate technique, it was compared with the independent univariate PLN models with respect to model inference, goodness-of-fit, identification of hazardous locations and precision of expected collision frequency.

The development of the MVPLN model is undertaken using the WinBUGS platform which facilitates computation of posterior distributions as well as providing a measure for model comparisons.

The estimates of the extra Poisson variation parameters were considerably smaller under MVPLN. As precision is inversely proportional to the variance of expected collision frequency, it is estimated that the MVPLN model is more than twice as precise as the univariate PLN models. The improvement in precision is due mainly to the fact that MVPLN accounts for the correlation between the latent variables (*PDO*) and (*I* + *F*). This correlation (ρ) was estimated at 0.758, which is highly significant. The results indicate that higher *PDO* rates are associated with higher *I* + *F* rates, as the collision likelihood for both types is likely to rise due to similar deficiencies in roadway design and/or other unobserved factors. This correlation was identified by other researchers in the literature.

In terms of the goodness-of-fit, the MVPLN provided a superior fit over the two univariate models as its (multivariate) DIC is much less than the sum of their (univariate) DICs. The differences between the DICs show a significant drop off of 41.6.

In term of model application, the paper introduces a new multivariate technique for the identification of hazardous locations. The new technique generalizes the univariate posterior probability of excess that has been commonly proposed and applied in the literature to fit the multivariate relationship between latent variables. The results showed that all the hazardous locations identified by the univariate models were identified by multivariate model (except one location under *PDO* PLN). The results also demonstrated that some hazardous locations could be overlooked if the analysis was restricted to the univariate models.

The results presented in this paper are based on a single dataset. Even though these results conform to those in the literature, further research with different datasets and infrastructures is required to confirm the paper's findings.

In this paper two severity levels were investigated (namely: *PDO* and *I* + *F*). However, different severity levels could be incorporated (e.g., fatality, severe injury, light injury, PDO, etc.) or different collisions types could be considered (e.g., angle, head-on, rear-end, sideswipe, etc.). The multivariate nature of such data sets is both logical and intuitive. Recent improvements in statistical tools allows for a more precise analysis which takes into account the correlations that exist among the different levels. These tools should be used to assist transportation engineers to better understand the relationships between the different modeling variables and techniques.

References

Aguero-Valverde, J., Jovanis, P.P., 2009. Bayesian multivariate Poisson log-normal models for crash severity modeling and site ranking. In: Presented at the 88th Annual Meeting of the Transportation Research Board.

Anastasopoulos, P.Ch., Mannering, F., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41 (1), 153–159.

Bedrick, E.J., Christensen, R., Johnson, W., 1996. A new perspective on priors for generalized linear models. Journal of the American Statistical Association 91, 1450–1460.

Brijs, T., Karlis, D., Van den Bossche, F., Wets, G., 2007. A Bayesian model for ranking hazardous road sites. Journal of the Royal Statistical Society: Series A (Statistics in Society) 170 (4), 1001–1017.

Brooks, S.P., Gelman, A., 1998. Alternative methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7, 434–455.

Cheng, W., Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accident Analysis and Prevention 37, 870–881.

Chib, S., Winkelmann, R., 2001. Markov chain Monte Carlo analysis of correlated count data. Journal of Business and Economic Statistics 19, 428–435.

Clayton, D., Kaldor, J., 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43 (3), 671–681.

Congdon, P., 2006. Bayesian Statistical Modeling, 2nd edition. Wiley, New York.

El-Basyouny, K., Sayed, T., 2006. Comparison of two negative binomial regression techniques in developing accident prediction models. Transportation Research Record 1950, 9–16.

Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., Thomsen, L.K., 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accident Analysis and Prevention 27 (1), 1–20.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall, London.
Hauer, E., 1997. Observational Before–After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Elsevier Science Ltd.
Hauer, E., Ng, J.C.N., Lovell, J., 1988. Estimation of safety at signalized intersections. Transportation Research Record 1185, 48–61.
Heydecker, B.G., Wu, J., 2001. Identification of sites for accident remedial work by Bayesian statistical methods: an example of uncertain inference. Advances in Engineering Software 32, 859–869.
Higle, J.L., Witkowski, J.M., 1988. Bayesian identification of hazardous sites. Transportation Research Record 1185, 24–35.
Hinde, J., Demetrio, C.G.B., 1998. Over-dispersion: model and estimation. Computational Statistics & Data Analysis 27 (2), 151–170.
Hughes, W., Eccles, K., Harwood, D., Potts, I., Hauer, E., 2005. Development of a Highway Safety Manual. Appendix C: Highway Safety Manual Prototype Chapter: Two-Lane Highways. NCHRP Web Document 62 (Project 17-18(4)). Washington, D.C. Available from http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w62.pdf.
Hydén, C., 1987. The Development of a Method for Traffic Safety Evaluation: The Swedish Traffic Conflicts Technique. Lund University, Sweden.
Karlis, D., Meligkotsidou, L., 2005. Multivariate Poisson regression with covariance structure. Statistics and Computing 15, 255–265.
Karlis, D., 2003. An EM algorithm for multivariate Poisson distribution and related models. Journal of Applied Statistics 30, 63–77.
Kim, H., Sun, D., Tsutakawa, R.K., 2002. Lognormal vs. gamma: extra variations. Biometrical Journal 44 (3), 305–323.
Kim, J.-K., Ulfarsson, G.F., Shankar, V.N., Kim, S., 2008. Age and pedestrian injury severity in motor-vehicle crashes: a heteroskedastic logit analysis. Accident Analysis and Prevention 40 (5), 1695–1702.
Kumara, S.S.P., Chin, H.C., 2006. Disaggregate models to examine signalized intersection crash frequencies. In: Presented at the 85th Annual Meeting of the Transportation Research Board, Washington, DC.
Ladron de Guevara, F., Washington, S., 2004. Forecasting crashes at the planning level. A simultaneous negative binomial crash model applied in Tucson, Arizona. Transportation Research Record 1897, 191–199.
Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-road accidents: an empirical analysis. Accident Analysis and Prevention 34 (2), 349–361.
Li, C.C., Lu, J.C., Park, J., Kim, K., Brinkley, P.A., Peterson, J.P., 1999. Multivariate zero-inflated Poisson models and their applications. Technometrics 41 (1), 29–38.
Lord, D., Persaud, B., 2000. Accident prediction models with and without trend: application of the generalized estimating equation. Transportation Research Record 1717, 102–108.
Lord, D., 2000. The prediction of accidents on digital networks: characteristics and issues related to the application of accident prediction models. Ph.D. Dissertation. Department of Civil Engineering, University of Toronto, Toronto, Ontario, Canada.
Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis and Prevention 38 (4), 751–766.
Lord, D., Guikema, S.D., Geedipally, S.R., 2008. Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis and Prevention 40, 1123–1134.
Lord, D., Park, Y.-J., 2008. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. Accident Analysis and Prevention 40 (4), 1441–1457.
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. Accident Analysis and Prevention 37 (1), 35–46.
Lord, D., Washington, S.P., Ivan, J.N., 2007. Further notes on the application of zero inflated models in highway safety. Accident Analysis and Prevention 39 (1), 53–57.
Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10, 325–337.
Ma, J., Kockelman, K.M., 2006. Bayesian multivariate Poisson regression for models of injury count by severity. Transportation Research Record 1950, 24–34.
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, 964–975.
Malyshkina, N.V., Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: an application to vehicle accident frequencies. Accident Analysis and Prevention 41, 217–226.
MATLAB Neural Network Toolbox 5, 2006. MathWorks, Inc., Natick, Massachusetts.
Maycock, G., Hall, R.D., 1984. Accidents at 4-arm roundabouts. TRRL Laboratory Report 1120. Transportation and Road Research Laboratory, Crowthorne, Bershire.
Miaou, S.P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes. Transportation Research Record 1840, 31–40.
Milton, J., Shankar, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention 40 (1), 260–266.
Miranda-Moreno, L.F., 2006. Statistical models and methods for identifying hazardous locations for safety improvements. Ph.D. Dissertation. Department of Civil Engineering, University of Waterloo, Waterloo, Ontario, Canada.
Miranda-Moreno, L.F., Fu, L., Saccomanno, F.F., Labbe, A., 2005. Alternative risk models for ranking locations for safety improvement. Transportation Research Record 1908, 1–8.
Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39, 459–468.
Park, E.S., Lord, D., 2007. Multivariate Poisson–lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, 1–6.
Persaud, B.N., Dzbik, L., 1993. Accident prediction models for freeways. Transportation Research Record 1401, 55–60.
Poch, M., Mannering, F.L., 1996. Negative binomial analysis of intersection-accident frequency. Journal of Transportation Engineering 122 (2), 105–113.
Qin, X., Ivan, J.N., Ravishankar, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis and Prevention 36 (2), 183–191.
SAS, 2002. Version 9 of the SAS System for Windows. SAS Institute Inc., Cary, NC.
Sawalha, Z., Sayed, T., 2006a. Transferability of accident prediction models. Journal of Safety Science 44 (3), 209–219.
Sawalha, Z., Sayed, T., 2006b. Traffic accidents modeling: some statistical issues. Canadian Journal of Civil Engineering 33 (9), 1115–1124.
Sayed, T., Abdelwahab, W., 1997. Using accident correctability to identify accident prone locations. Journal of Transportation Engineering, ASCE 123 (2), 107–113.
Sayed, T., de Leur, P., 2007. Evaluation of Edmonton's intersections safety camera programs. Transportation Research Record 2009, 37–45.
Sayed, T., Zein, S., 1999. Traffic conflict standards for intersections. Transportation Planning and Technology 22, 309–323.
Schluter, P.J., Deely, J.J., Nicholson, A.J., 1997. Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. The Statistician 46 (3), 293–316.
Shankar, V.N., Chayanan, S., Sittikariya, Shyu, M.-B., Juvva, N.K., Milton, J.C., 2004. Marginal impacts of design, traffic, weather, and related interactions on roadside crashes. Transportation Research Record 1897, 156–163.
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.
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 B 64, 1–34.
Spiegelhalter, D.J., Thomas, A., Best, N.G., 1996. Computation on Bayesian graphical models. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 5. Oxford University Press, Oxford, pp. 407–425.
Tsionas, E.G., 2001. Bayesian multivariate Poisson regression. Communications in Statistics-Theory and Methods 30, 243–255.
Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. Austrian Journal of Statistics 31 (3), 221–229.
Warton, D.I., 2005. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16 (2), 275–289.
Wood, G.R., 2002. Generalized linear accident models and goodness of fit testing. Accident Analysis and Prevention 34 (4), 417–427.
Yamamoto, T., Hashiji, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accident Analysis and Prevention 40 (4), 1320–1329.
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.