Accident Analysis and Prevention 43 (2011) 1160–1166

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A joint-probability approach to crash prediction models

Xin Pei[1], S.C. Wong[2], N.N. Sze*

Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong

ARTICLE INFO

Article history:
Received 14 September 2010
Received in revised form
18 December 2010
Accepted 22 December 2010

Keywords:
Crash frequency
Crash severity
Joint probability
Full Bayesian method
Markov chain Monte Carlo (MCMC) approach

ABSTRACT

Many road safety researchers have used crash prediction models, such as Poisson and negative binomial regression models, to investigate the associations between crash occurrence and explanatory factors. Typically, they have attempted to separately model the crash frequencies of different severity levels. However, this method may suffer from serious correlations between the model estimates among different levels of crash severity. Despite efforts to improve the statistical fit of crash prediction models by modifying the data structure and model estimation method, little work has addressed the appropriate interpretation of the effects of explanatory factors on crash occurrence among different levels of crash severity. In this paper, a joint probability model is developed to integrate the predictions of both crash occurrence and crash severity into a single framework. For instance, the Markov chain Monte Carlo (MCMC) approach full Bayesian method is applied to estimate the effects of explanatory factors. As an illustration of the appropriateness of the proposed joint probability model, a case study is conducted on crash risk at signalized intersections in Hong Kong. The results of the case study indicate that the proposed model demonstrates a good statistical fit and provides an appropriate analysis of the influences of explanatory factors.

1. Introduction

It has been predicted that road crashes will become the fifth leading cause of death worldwide by 2030 (World Health Organization, 2009). Road fatalities and injuries result in loss of life and property, and decreases in quality of life. A better understanding of the factors associated with crashes, injuries, and death is critical to enhancing the safety performance of road traffic systems.

The number of road crashes occurring at an entity per unit of time is non-negative, discrete, and random. It is thus not appropriate to model crash occurrence by the standard least squares method. Several count data models, including the Poisson and negative binomial regression models, have been developed and applied to investigate the relationship between crash occurrence and the explanatory factors associated with various traffic circumstances (Miaou and Lum, 1993; Miaou, 1994; Poch and Mannering, 1996; Milton and Mannering, 1998).

Crash severity is of particular interest to road safety engineers and researchers. Crashes can be classified according to the most severe injury caused, such as fatality, disabling injury, evident injury, possible injury, and property damage only. In Hong Kong, crash severity is divided into three categories – fatal, severe, and slight – according to the injury severity of the most seriously injured person in the crash. The most common approach is to first model the total crash frequency and then the conditional crash severity. However, detailed crash-specific information, such as driver demographics, driving experience, casualty particulars, vehicle attributes, and the use of restraint measures, are required for a robust analysis of conditional crash severity (Carson and Mannering, 2001; Lee and Mannering, 2002).

To address the issues of severity and frequency together, some researchers have estimated separate and independent crash frequency models for each discrete severity outcome (i.e., a frequency model of fatalities and a frequency model of disabling injuries). However, there is always the possibility of a significant correlation among the counts for the different levels of severity that will interfere with the model estimates (Lord and Mannering, 2010). To deal with this problem, researchers have devised various advanced modeling approaches, including multi-level hierarchical structures (Kim et al., 2007; Huang and Abdel-Aty, 2010), simultaneous equations (Kim and Washington, 2006; Ye et al., 2009), and multivariate analysis (Ma and Kockelman, 2006; Park and Lord, 2007; Ma et al., 2008).

For instance, the multivariate generalized linear model approach explicitly considers the correlation among different levels of crash severity. This improves the association measures by analyzing crashes of different severity levels simultaneously using

* Corresponding author. Tel.: +852 2859 2662; fax: +852 2517 0124.
E-mail addresses: peix@hkusua.hku.hk (X. Pei), hhecwsc@hkucc.hku.hk (S.C. Wong), nnsze@graduate.hku.hk (N.N. Sze).
[1] Tel.: +852 2859 2662; fax: +852 2517 0124.
[2] Tel.: +852 2859 1964; fax: +852 2559 5337.

0001-4575/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2010.12.026

the same set of risk factors. This strengthens the weak inferences regarding risk factors obtained from individual crash frequency analyses (Bailey and Hewson, 2004; Bijleveld, 2005; Miaou and Song, 2005; Song et al., 2006). This, in turn, resolves the correlation problems associated with isolated crash prediction models and improves the prediction performance by achieving a superior statistical fit. In this study, we establish a joint probability model that analyzes crash occurrence and crash severity simultaneously. The proposed joint probability model is less complicated than multivariate count data models in terms of the correlation structure between the frequency of crashes of different levels of severity. The proposed model is also less data intensive than conditional crash severity models. This allows the prediction capability of the model to be improved by the use of limited unconditional crash data.

In the following sections, we first describe the framework and mechanism of the proposed joint probability model. We then demonstrate the application of the model in an analysis of the frequency of crashes of two severity levels at signalized intersections in Hong Kong. The risk factors contributing to the two crash types are also assessed. We then assess and evaluate the results of the illustrative example to reach some conclusions on the appropriateness and limitations of the proposed model and suggest some ways forward.

2. Methodology

2.1. Modeling framework

In light of a possible correlation among crash frequencies of different severity levels, we establish a joint probability model, analyzing the influences of crash characteristics on crash occurrence and crash severity simultaneously.

2.1.1. Probability function for crash occurrence

For the crash occurrence, considering the discrete, non-negative and random nature of crash frequency, a Poisson regression model is established with the probability function defined as (Washington et al., 2003)

$$P(y_i) = \frac{\exp(-\lambda_i)\lambda_i{}^{y_i}}{y_i!} \tag{1}$$

where λ_i is the expected number of crash at the ith entity and y_i is the observed number of crash at the ith entity respectively.

The relationship between λ_i and explanatory factors is governed by a log-linear function defined as

$$\lambda_i = \exp(\theta \mathbf{X}_i) \tag{2}$$

where $\mathbf{X}_i$ denotes the vector of explanatory factors and $\boldsymbol{\theta}$ the vector of corresponding coefficients.

The basic assumption of Poisson's distribution is that the mean of the count process equals its variance. When the data is subjected to over-dispersion (that is, the variance of the count is significantly greater than its mean), negative binomial regression should be adopted. Cameron and Trevedi (1990) suggested a test to investigate the existence of over-dispersion in the data, of which the null and alternative hypotheses are specified as

$$\begin{cases} \mathrm{H}_0 : \mathrm{Var}(y_i) = E(y_i) \\ \mathrm{H}_1 : \mathrm{Var}(y_i) = E(y_i) + \alpha g\,[E(y_i)] \end{cases} \tag{3}$$

where $g[E(y_i)]$ is given by either $g[E(y_i)] = E(y_i)$ for NB-1 model or $g[E(y_i)] = E(y_i)^2$ for NB-2 model respectively.

Above is the traditional parameterization for the negative binomial regression model (Hilbe, 2007). Then, a simple linear regression model on z can be defined as

$$\mathbf{z} = \eta \mathbf{w} \tag{4}$$

where $z_i = (y_i - E(y_i))^2 - y_i/\sqrt{2}E(y_i)$ and $w_i = g[E(y_i)]/\sqrt{2}E(y_i)$.

If η is statistically significant in either case (NB-1 or NB-2 model), then H_0 is rejected in favor of H_1. Hence, the data is said to be over-dispersed, and the negative binomial regression model should be applied.

The negative binomial model can be governed by a log-linear function defined as

$$\lambda_i = \exp(\theta \mathbf{X}_i + \varepsilon_i) \tag{5}$$

where $\exp(\varepsilon_i)$ is a gamma-distributed error with mean and variance of 1 and α^2 respectively. α is the over-dispersion parameter.

Hence, the probability function of the negative binomial model can be defined by

$$P(y_i) = \frac{\Gamma[(1/\alpha) + y_i]}{\Gamma(1/\alpha) y_i!} \left(\frac{1/\alpha}{(1/\alpha) + \lambda_i}\right)^{1/\alpha} \left(\frac{\lambda_i}{(1/\alpha) + \lambda_i}\right)^{y_i} \tag{6}$$

2.1.2. Probability function for crash severity

For crash severity upon occurrence, both the hierarchical binomial-logistic approach and the hierarchical truncated Poisson's approach are considered to model the relationship between crash severity and explanatory factors.

Theoretically, each crash occurrence can be considered as a Bernoulli trial. The crash outcome in terms of severity level should then follow a binomial distribution. Here, a binomial-logistic model is established to model the influence of explanatory factors on crash severity. For instance, number of crash k^l of severity level l follows a binomial distribution defined as

$$k_i^l \sim \mathrm{Binomial}(\pi_i^l, y_i) \tag{7}$$

where y_i is the observed number of crashes at the ith entity given in Eq. (1) and π_i^l denotes the probability of crash severity level l respectively.

To model the association between binomial probability and explanatory factors, we can establish a logit function defined as

$$\mathrm{logit}(\pi_i^l) = \log\left(\frac{\pi_i^l}{1 - \pi_i^l}\right) = \beta^l \mathbf{X}_i \tag{8}$$

where β^l denotes the effect of explanatory factors on the crash outcome of severity level l.

The probability function of k_i^l crashes of severity level l conditional on y_i total crashes can thus be defined as

$$P(k_i^l | y_i) = \binom{y_i}{k_i^l} (\pi_i^l)^{k_i^l} (1 - \pi_i^l)^{y_i - k_i^l} \tag{9}$$

Hence, we have

$$P(k_i^l | y_i) = \binom{y_i}{k_i^l} \left[\frac{\exp(\beta^l \mathbf{X}_i)}{(1 + \exp(\beta^l \mathbf{X}_i))}\right]^{k_i^l} [(1 + \exp(\beta^l \mathbf{X}_i))^{-1}]^{y_i - k_i^l}$$

Alternatively, the number of crashes k^l of severity level l can be considered as truncated count data, given which the count is right-truncated by the total number of crashes y_i. Therefore, the conditional probability of k^l crashes can be defined as follows:

$$P(k_i^l | y_i) = [\lambda_i{}^{y_i}/y_i!] \Big/ \left[\sum_{m_i=0}^{y_i} \lambda_i^{m_i}/m_i!\right]$$
$$P(k_i^l | y_i) = [\exp(\beta^l \mathbf{X}_i y_i)/y_i!] \Big/ [\sum_{m_i=0}^{y_i} \exp(\beta^l \mathbf{X}_i m_i)/m_i!] \tag{10}$$

Interested readers can refer to Washington et al. (2003) for a detailed formulation of the truncated Poisson's model.

2.1.3. Joint probability

Integrating the probability function for crash occurrence (Eqs. (1) and (6)) and crash severity (Eqs. (9) and (10)), the joint probability function of having y_i total crashes and k_i^l crashes of severity level l can be defined as

$$P(y_i, k_i^l) = P(y_i) \times P(k_i^l|y_i) = f(\theta \mathbf{X}_i) \times f(\beta^l \mathbf{X}_i) \quad (11)$$

where $\boldsymbol{\theta}$ denotes the effects of explanatory factors on crash occurrence and β denotes the effects of explanatory factors on crash severity.

2.2. Model estimation

2.2.1. Full Bayesian's approach

A simulation-based full Bayesian's approach will be employed to model the joint probability of crash occurrence and crash severity. The Bayesian approach is a modeling technique that considers and characterizes all unknown parameters as random variables under a prior distribution. Instead of giving point estimations in the classical maximum likelihood approach, the Bayesian approach can summarize the results by providing a posterior (probability) distribution of the parameters in the proposed model. For instance, estimates of means, standard deviation, and quartiles of parameters of explanatory factors can be determined.

Based on the specification of Bayes' theorem and the integration of the prior distribution of parameters and likelihood function, the posterior distribution of parameters can be estimated by the function defined as

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} \propto f(y|\theta)f(\theta) \quad (12)$$

where $\boldsymbol{y}$ is observed outcome and θ is parameter estimate.

Therefore, the likelihood function can be defined as

$$f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta) \quad (13)$$

Appropriate specification of prior distribution $f(\theta)$ is critical in Bayesian's inference. For instance, it is essential to have prior information on and good understanding of interested variables. If prior information is not available, we can set up a diffuse prior distribution as

$$\theta \sim \text{Normal}(0, \sigma^2 I_m) \quad (14)$$

where σ^2 is very large and I is an identity matrix of m dimensions, with which m equals to number of parameters.

2.2.2. Markov chain Monte Carlo (MCMC) simulation

Markov chain Monte Carlo (MCMC) simulation is a 'golden' approach for Bayesian's inference. For instance, the MCMC method draws the sample from the prior distribution of unknown parameters in an iterative procedure. Such a process should repeat until the distribution of parameter estimates converges and gives the posterior distribution.

The Metropolis–Hastings algorithm and Gibbs sampling are two common sampling approaches in MCMC method.

In the simulation process of the Metropolis–Hastings algorithm, an initial value of a parameter is first given, say $\theta^{(0)}$. Then, in each iteration t, a candidate value $\theta^{(t)}$ will be given by the function of $\theta^{(t-1)}$, where such a candidate is only accepted at α level of significance. Hence, a Markov chain $\{\theta^{(0)}, \theta^{(1)}, ..., \theta^{(t-1)}, \theta^{(t)}, ...\}$ of a random sample of parameter estimate can be generated. When the parameter estimate converges, summary statistics of the posterior distribution of concerned parameter can be determined.

Gibbs sampling can be regarded as a special case of the Metropolis–Hastings sampling algorithm, in which a candidate $\theta^{(t)}$ is always accepted, i.e., α equal to one. In addition, the Gibbs sampling approach is capable of drawing a sample from the distribution of parameters other than that of the concerned variable, ensuring that the estimates will converge more easily. This has made the popular Gibbs sampling a 'golden standard'. Yet Gibbs sampling might not be suitable when the parameter is of a sophisticated form of distribution (Ntzoufras, 2009).

To evaluate the convergence of the estimates in the iterative procedure, a target posterior distribution needs to have been established. Practical ways for the evaluation of convergence include estimating the Markov chain error, plotting the autocorrelation distributions, plotting the generated sample values, and computing Gelman–Rubin statistics (Spiegelhalter et al., 2003).

2.2.3. Goodness-of-fit assessment

To evaluate the statistical fit of the proposed model, an information criterion is applied. Information criteria such as Akaike' s information criterion (AIC) and the Bayesian information criterion (BIC) are superior to the traditional likelihood ratio test statistic because they take into account the effect of the number of parameters and sample size.

In the full Bayesian model, the deviance information criterion (DIC) has been adapted to evaluate statistical fit (Spiegelhalter et al., 2002). In particular, the DIC can be estimated with a function defined as

$$DIC = D(\bar{\theta}) + 2p_D = \overline{D(\theta)} + p_D \quad (15)$$

where $D(\bar{\theta})$ is the usual deviance evaluated at the posterior means of parameter $\boldsymbol{\theta}$, $\overline{D(\theta)}$ is the posterior mean of deviance, and p_D is the effective number of parameters. Usual deviance can be determined by a function given as $D(\bar{\theta}) = -2 \log f(y|\bar{\theta})$.

As in the use of the AIC and BIC, the lower the value of the DIC, the better the statistical fit of the model.

The validity of the model can be tested by a comparison of the observed data with the replicated data, which is simulated from the posterior predictive distribution as the predictive data. Chi-square can be used as test statistics defined as

$$\chi^2(\boldsymbol{y}, \theta) = \sum_{i=1}^{n} \frac{[y_i - E(y_i|\theta)]^2}{\text{Var}(y_i|\theta)}, \quad (16)$$

where y_i denotes either the observed or the replicated response. In each iteration t of MCMC simulation, the difference between $\chi^2(y^{\text{rep}}, \theta^{(t)})$ and $\chi^2(y, \theta^{(t)})$ should be monitored, as should the corresponding posterior predictive p-value. A model does not fit the data well if p-values are close to 0 or 1, indicating that the predictive data are far from the observed data. (Gelman et al., 2004).

3. Data and model specification

3.1. Illustrative example

As an illustration of the appropriateness of the proposed joint probability model, we conducted a case study to model the crash frequencies of two classes in terms of crash severity level, i.e., total crashes, and killed and seriously injured (KSI) crashes, at the signalized intersections in Hong Kong. In a previous paper, the association measures for crash frequencies of different severity levels were dealt with separately (Wong et al., 2007). In this study, we revisit the problem, using the proposed joint probability model, with the

Table 1
Summary statistic of crash frequencies and factors considered.

	Range	Mean	S.D.
	Number of observation = 262		
Total crash	(Min: 0; Max: 24)	1.66	3.28
Killed and seriously injured (KSI) crash	(Min: 0; Max: 5)	0.32	0.72
Log (AADT)	(Min: 6.81; Max: 11.71)	9.91	0.76
Number of approaches	(Min: 2; Max: 5)	3.56	0.61
Average lane width	(Min: 2.7; Max: 5.5)	3.38	0.38
Reciprocal of the average turning radius	(Min: 0; Max: 0.2)	0.08	0.03
Proportion of commercial vehicles	(Min: 0.00; Max: 0.66)	0.19	0.10
Number of signal stages	(Min: 2; Max: 7)	3.07	0.90
Cycle time	(Min: 44; Max: 140)	92.80	19.22
Number of pedestrian crossing streams	(Min: 0; Max: 8)	3.50	2.09
Presence of tram stops	(1: Yes; 0: No)	0.07	
Presence of LRT stops	(1: Yes; 0: No)	0.04	
Presence of right turning pocket	(1: Yes; 0: No)	0.09	
Kowloon	(1: Yes; 0: No)	0.42	

MCMC approach full Bayesian's method. The effects of crash characteristics – traffic pattern, road environment, traffic control, and driver-related factors – on crash occurrence and crash severity are evaluated.

In this illustrative example, traffic accident data was obtained from the Traffic Accident Database System (TRADS), maintained by the Hong Kong Police Force and the Transport Department. In TRADS, crashes are divided into three categories, according to crash severity levels, namely fatal, serious, and slight. The fatal and serious crashes are combined as killed and seriously injured (KSI) crashes in the subsequent analysis. In addition, information about traffic volume, geometric design, road environment, and signal phasing are obtained from a comprehensive set of traffic impact assessment (TIA) reports for 2002 and 2003. Crash information of 262 signalized junctions, which account for 15.7% (out of 1660) of the signalized intersections in Hong Kong, was obtained. Interested readers can refer to the previous paper (Wong et al., 2007) for a detailed description of the crash data used in this study.

Drawing on the results of correlation analysis, twelve neutral, independent variables are selected for subsequent crash prediction models. The variables selected include traffic volume, number of approaches, average lane width, reciprocal of the average turning radius, presence of tram stops, presence of LRT stops, proportion of commercial vehicles, number of signal stages, cycle time, number of pedestrian crossing streams, presence of right turning pocket, and Kowloon area. Table 1 presents the summary statistics of dependent variables (i.e., total crashes and KSI crashes) and independent variables.

3.2. Model specification

The proposed joint probability model is applied to evaluate the effects of explanatory factors on crash occurrence and crash severity at signalized intersections in Hong Kong, using the MCMC approach full Bayesian's method.

For crash occurrence, count data models are applied to measure the effects of related factors. Based on the results of the over-dispersion test, the crash data is subject to significant over-dispersion at the 1% level of significance, both in NB-1 (t-statistic = 3.09) and NB-2 (t-statistic = 2.91) models. Therefore, the negative binomial regression model should be applied to model the crash occurrence. The NB-2 model is applied in the illustrative example.

Crash severity is divided into two classes: KSI crashes and slight crashes. The proposed joint probability model is then applied to the number of KSI crashes.

Using the hierarchical binomial-logistic model given in Eqs. (7)–(9), the number of KSI crashes k_i^{KSI} should follow a distribution as

$$k_i^{\mathrm{KSI}} \sim \mathrm{Binomial}(\pi_i^{\mathrm{KSI}}, y_i) \tag{17}$$

where y_i is the total crash at intersection i, and π_i^{KSI} is the probability of KSI crashes.

Then, the logit function can be defined as

$$\mathrm{logit}(\pi_i^{\mathrm{KSI}}) = \beta^{\mathrm{KSI}}\mathbf{X}_i \tag{18}$$

where β^{KSI} is the vector of parameters representing the effects of explanatory factors on crash severity.

Using the right-truncated count data model defined in Eq. (10), the number of KSI crashes k_i^{KSI} should follow a distribution as

$$k_i^{\mathrm{KSI}} \sim \mathrm{Truncated\_Poisson}(\lambda_i^{\mathrm{KSI}}, y_i) \tag{19}$$

where λ_i^{KSI} is the predicted number of KSI crash and is defined as

$$\lambda_i^{\mathrm{KSI}} = \exp(\beta^{\mathrm{KSI}}\mathbf{X}_i)$$

where β^{KSI} is the vector of parameters representing the effects of explanatory factors on KSI crash.

Consequently, a joint probability model is established to predict the number of total crashes, y_i and KSI crashes, k_i^{KSI} at the same time. For instance, two joint probability models are developed, one by the negative binomial–binomial logistic (NBBL) approach (shown in Eq. (20)) and one by the negative binomial-truncated Poisson's (NBTP) approach (shown in Eq. (21)), respectively.

$$\begin{aligned} P(y_i, k_i^{\mathrm{KSI}}) = P(y_i) \times P(k_i^{\mathrm{KSI}}|y_i) &= \frac{\Gamma[(1/\alpha)+y_i}{\Gamma(1/\alpha)y_i!}\left(\frac{1/\alpha}{(1/\alpha)+\exp(\theta\mathbf{X}_i)}\right)^{1/\alpha} \\ &\times \left(\frac{\exp(\theta\mathbf{X}_i)}{(1/\alpha)+\exp(\theta\mathbf{X}_i)}\right)^{y_i} \\ &\times \binom{y_i}{k_i^{\mathrm{KSI}}} [\exp(k_i^{\mathrm{KSI}}\beta^{\mathrm{KSI}}\mathbf{X}_i)][1+\exp(\beta^{\mathrm{KSI}}\mathbf{X}_i)]^{-y_i} \end{aligned} \tag{20}$$

$$\begin{aligned} P(y_i, k_i^{\mathrm{KSI}}) = P(y_i) \times P(k_i^{\mathrm{KSI}}|y_i) &= \frac{\Gamma[(1/\alpha)+y_i}{\Gamma(1/\alpha)y_i!}\left(\frac{1/\alpha}{(1/\alpha)+\exp(\theta\mathbf{X}_i)}\right)^{1/\alpha} \\ &\times \left(\frac{\exp(\theta\mathbf{X}_i)}{(1/\alpha)+\exp(\theta\mathbf{X}_i)}\right)^{y_i} \times \frac{[\exp(\beta^{l}\mathbf{X}_i y_i)/y_i!]}{\left[\sum_{m_i=0}^{y_i} exp(\beta^{l}\mathbf{X}_i m_i)/m_i!\right]} \end{aligned} \tag{21}$$

Table 2
Results of joint probability model.

	Crash occurrence		Crash severity	
	Mean	95% CI	Mean	95% CI
Number of observation = 262				
Explanatory factors				
Constant	−7.12*	(−11.37, −3.13)	10.09*	(3.99, 15.21)
Log(AADT)	0.64*	(0.39, 0.92)	−0.84*	(−1.19, −0.46)
Kowloon area	0.84*	(0.42, 1.24)	−0.34	(−0.93, 0.27)
Presence of tram stops	1.18*	(0.40, 1.99)	0.54	(−0.51, 1.61)
Proportion of commercial vehicles	2.82*	(0.78, 4.89)	0.67	(−2.07, 3.35)
Average lane width	−0.54	(−1.18, 0.08)	−1.34*	(−2.30, −0.33)
Number of approaches	0.16	(−0.27, 0.63)	−0.004	(−0.69, 0.71)
Reciprocal of the average turning radius	6.37	(−1.08, 14.05)	0.44	(−9.50, 11.02)
Presence of LRT stops	−0.99	(−2.51, 0.45)	1.32	(−1.10, 3.72)
Number of signal stages	−0.25	(−0.54, 0.04)	0.32	(−0.10, 0.74)
Cycle time	0.01	(−0.001, 0.02)	0.003	(−0.02, 0.02)
Number of pedestrian streams	0.07	(−0.03, 0.18)	0.10	(−0.05, 0.26)
Presence of right turning pocket	−0.01	(−0.67, 0.69)	−0.77	(−1.87, 0.23)
Over-dispersion parameter				
$1/\alpha$	0.78*	(0.54, 1.11)		
Goodness-of-fit				
DIC	1078.32			
Posterior predictive *p*-value	0.82			

* Statistically significant at the 5% level.

4. Results

In this study, the MCMC approach full Bayesian's method is employed to estimate the parameters in the proposed joint probability models given in Eqs. (20) and (21), respectively. For instance, three chains of 30,000 iterations are adopted in each of the simulation processes concerned. All predicted values of concerned coefficients fall in the range that does not produce strong periodicities and tendencies, as indicated by the trace plots. Moreover, using the Gelman–Rubin convergence diagnostic, results show that both the inter-sample and intra-sample variability are stable and are close to one. In other words, the model converges.

The performance of the NBBL model (DIC = 1078.32) is superior to that of the NBTP model (DIC = 1086.00). Besides, the χ^2 statistic can be used in goodness-of-fit assessment. Again, the results of the χ^2 test indicated that the NBBL model (posterior predictive *p*-value = 0.82) is superior to the NBTP model (posterior predictive *p*-value = 0.95), at the 5% level of significance. Therefore, NBBL model is adopted.

Table 2 illustrates the results of the MCMC approach full Bayesian's method for the effects of explanatory factors on crash occurrence and crash severity, respectively.

As shown in Table 2, traffic volume [mean = 0.64; 95% CI = (0.39, 0.92)], geographical area [mean 0.84, 95% CI (0.42, 1.24)], presence of tram stops [mean 1.18, 95% CI (0.40, 1.99)], and proportion of commercial vehicles [mean 2.82, 95% CI (0.78, 4.89)] determine the crash occurrence, all at the 5% level of significance.

Traffic volume [mean = −0.84; 95% CI = (−1.19, −0.46)] and average lane width [Mean −1.34, 95% CI (−2.30, −0.33)] determine the crash severity, both at the 5% level of significance, as also shown in Table 2.

5. Discussion

Using the results of parameter estimates of the proposed joint probability model, in Table 3 we categorize the factors into four groups: those influencing both crash occurrence and crash severity [Quadrant 1]; those influencing crash occurrence only [Quadrant 2]; those influencing crash severity only [Quadrant 3]; and no evidence that associations with crash occurrence and crash severity can be established [Quadrant 4]. It is also worth exploring the occasions when the directions of the effects of related factors vary in crash occurrence and crash severity.

The influences of explanatory factors on crash occurrence and crash severity are now examined, with respect to the categories listed above.

Quadrant 1: Significant factors in both crash occurrence and crash severity

Traffic volume significantly determined both crash occurrence and crash severity, but the directions of the effects among the two were different. For instance, the frequency of total crashes increases with the traffic volume, while the proportion of KSI crashes in total crashes decreases with the traffic volume.

Table 3
Significant factors contributing to crash occurrence and crash severity, respectively.

		Crash occurrence	
		Significant	Not significant
Crash severity	Significant	**Quadrant 1** Log(AADT)	**Quadrant 3** Average lane width
	Not significant	**Quadrant 2** Kowloon area Presence of tram stops Proportion of commercial vehicles	**Quadrant 4** Number of approaches Average turning radius Presence of LRT stops Number of signal stages Cycle time Number of pedestrian streams Presence of right turning pocket

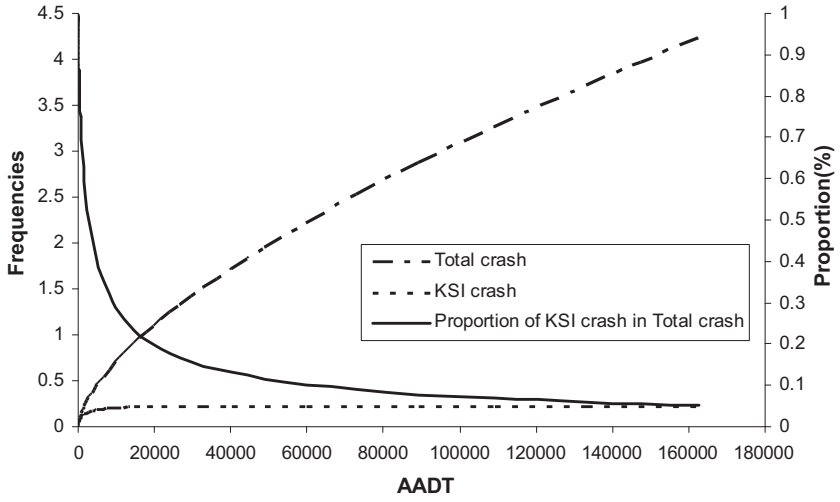

Fig. 1. Relationship between crash frequencies and AADT.

Fig. 1 presents the relationship between traffic volume and frequencies of (i) total crashes, (ii) KSI crashes, and (iii) proportion of KSI crashes in total crashes. As expected, the frequencies of total crashes and KSI crashes increased with traffic volume, as the opportunities for traffic conflict likewise increased. However, it is apparent that the increasing rate of the frequency of KSI crashes is lower than that of total crashes. The frequency of KSI crashes reaches the maximum value and starts to decrease when traffic volume increases beyond 60,000 (vehicle per day). Therefore, the proportion of KSI crashes (in total crashes) should decrease with traffic volume, perhaps through the reduced risk of fatality and severe injury following from reduced traffic speed and thus energy released in collisions under heavy traffic conditions. This is consistent with the findings of several previous studies (Mountain et al., 1996; Qin et al., 2004, 2006). In particular, results have indicated that the frequency of fatal crashes increases at a lower rate than that of total crashes as traffic volume increases (Fridstrøm et al., 1995).

Quadrant 2 Significant factors to crash occurrence only

For geographical characteristics, the risk of crash occurrence at signalized intersections in Kowloon is higher than that in other geographical areas. This might be because of the distinctive land use in Kowloon. Since the early 20th century, most of Kowloon has been heavily developed, resulting in a dense urban area of both residential and commercial use. Roads in the Kowloon area are narrow and highly congested, with frequent off-road commercial activities. Therefore, the risk of traffic conflict and crash occurrence might have become high. Previous studies have consistently found a link between crash risk and land use characteristics (Kim and Yamashita, 2002; El-Basyouny and Sayed, 2009).

In terms of the traffic patterns, the presence of tram stops increases the risk of crash occurrence at signalized intersections. Frequent pedestrian activity occurs around tram stops, with potential vehicle–pedestrian conflicts. In addition, variations in the vehicular speed of mixed traffic (trams, public transport vehicles, and private vehicles) might increase the risk of traffic conflicts. Trams share their lanes with other motor vehicles at numerous busy intersections, which might further increase the risk of traffic conflict and vehicle collision. In particular, most of the tram passengers are vulnerable road users, such as the elderly and school children, who take advantage of lower fare of tram service. Previous studies have addressed the problem of crash risk near tram stops in Gothenburg, Sweden (Hedelin et al., 1996), Sheffield, UK (Cameron et al., 2001) and Austria (Unger et al., 2002).

In addition, crash risk at signalized intersection increases in proportion with the number of commercial vehicles. The frequent drop-off and pick-up activities of commercial vehicles, such as public buses and trucks, might increase the opportunity for traffic conflicts, especially in congested urban area like Hong Kong. The visibility of commercial drivers and other road users is reduced due to vehicle dimensions, and, therefore, crash risk might increase. Moreover, stress, fatigue, and the aggressive driving of some reckless commercial drivers would also increase crash risk (Boufous and Williamson, 2006; Brodie et al., 2009).

Quadrant 3 Significant factors to crash severity only

In terms of geometric design, the results indicated that the risk of severe injury crashes decreased with the average lane width. This might be attributable to the increase in the opportunity for defensive maneuvers upon crashes, due to the increase in vehicle–vehicle and vehicle–road infrastructure separation on a wider road. Controversial arguments have been made in other studies on crash occurrence on highway segments (Shankar et al., 1997; Milton and Mannering, 1998; Rengarasu et al., 2009; Gross et al., 2009). The link between lane width and crash severity at signalized intersections has only rarely been considered.

6. Conclusion

In response to the problem of correlations in analyses of the frequency of crashes of different types when traditional isolated crash prediction models are used, we have developed a joint probability model to simultaneously evaluate the effects of explanatory factors on crash occurrence and crash severity for situations in which detailed crash-specific information are not available. An illustrative example using crash data from signalized intersections in Hong Kong is presented. The MCMC approach full Bayesian's method is employed, and the results indicate that the proposed model has a good statistical fit. The risk factors for crash occurrence and different levels of crash severity at a signalized intersection are evaluated. The results of the parameter estimates are appropriate in terms of the physical meaning of the associations for total and KSI crashes.

This study indicates that the proposed joint probability model is suitable for use as a crash prediction model for signalized intersections. Future research could extend the application of the model to other road entities and to other crash types, such as collision characteristics. Only the binary-severity case has been assessed with the proposed model as an illustrative example, but it would also be worth exploring whether the model could be extended to the multinomial-severity case.

Acknowledgements

We thank the two anonymous referees for the helpful suggestions and critical comments on an earlier version of the paper. The work described in this paper was partially supported by a Research Postgraduate Studentship, an Outstanding Researcher Award, the Engineering Postdoctoral Fellowship Programme of the University of Hong Kong, and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU7176/07E).

References

Bailey, T.C., Hewson, P.J., 2004. Simultaneous modeling of multiple traffic safety performance indicators by using a multivariate generalized linear mixed model. Journal of Royal Statistical Society Series A 167 (3), 501–517.
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, 591–600.
Boufous, S., Williamson, A., 2006. Work-related traffic crashes: a record linkage study. Accident Analysis and Prevention 38 (1), 14–21.
Brodie, L., Lyndal, B., Elias, I.G., 2009. Heavy vehicle driver fatalities: learning's from fatal road crash investigations in Victoria. Accident Analysis and Prevention 41 (3), 557–564.
Cameron, A., Trevedi, P., 1990. Regression based tests for overdispersion in the Poisson model. Journal of Econometrics 46, 347–364.
Cameron, I.C., Harris, N.J., Kehoe, N.J.S., 2001. Tram-related injuries in Sheffield. Injury 32 (4), 275–277.
Carson, J., Mannering, F., 2001. The effect of ice warning signs on accident frequencies and severities. Accident Analysis and Prevention 33, 99–109.
El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis and Prevention 41 (5), 1118–1123.
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.
Gelman, A., Carlin, John B., Stern, Hal S., Rubin, D.B., 2004. Bayesian Data Analysis, 2nd ed. Chapman & Hall/CRC.
Gross, F., Jovanis, P., Eccles, K., 2009. Safety effectiveness of lane and shoulder width combinations on rural, two-lane, undivided roads. Transportation Research Recordings 2103, 42–49.
Hedelin, A., Björnstig, U., Brismar, B., 1996. Trams—a risk factor for pedestrians. Accident Analysis and Prevention 28 (6), 733–738.
Hilbe, J.M., 2007. Negative Binomial Regression. Cambridge University Press, Cambridge.
Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accident Analysis and Prevention 42 (6), 1556–1565.
Kim, D.G., Washington, S.P., 2006. The significance of endogeneity problems in crash models: an examination of left-turn lanes in intersection crash models. Accident Analysis and Prevention 38, 1094–1100.
Kim, D.G., Lee, Y., Washington, S.P., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. Accident Analysis and Prevention 39, 125–134.
Kim, K., Yamashita, E., 2002. Motor vehicle crashes and land use—empirical analysis from Hawaii. Transportation Research Recording 1784, 73–79.
Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Analysis and Prevention 34, 149–161.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research 44A, 291–305.
Ma, J., Kockelman, K.M., 2006. Bayesian multivariate Poisson regression for models 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.
Miaou, S.P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accident Analysis and Prevention 25, 689–709.
Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention 26, 471–482.
Miaou, S.P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion and spatial dependence. Accident Analysis and Prevention 37, 699–720.
Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25, 395–413.
Mountain, L., Fawaz, B., Jarrett, D., 1996. Accident prediction models for roads with minor junctions. Accident Analysis and Prevention 28 (6), 695–707.
Ntzoufras, I., 2009. Bayesian Modeling Using WinBUGS. Wiley.
Park, E.S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, 1–6.
Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection accident frequencies. Journal of Transportation Engineering 122, 391–401.
Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accident Analysis and Prevention 36 (2), 183–191.
Qin, X., Ivan, John N., Ravishanker, N., Liu, J., Tepas, D., 2006. Bayesian estimation of hourly exposure functions by crash type and time of day. Accident Analysis and Prevention 38 (6), 1071–1080.
Rengarasu, T., Hagiwara, T., Hirasawa, M., 2009. Effects of road geometry and cross-section variables on traffic accidents. Transportation Research Recording 2102, 34–42.
Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accident Analysis and Prevention 29 (6), 829–837.
Song, J.J., Ghosph, 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. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64 (4), 583–639.
Spiegelhalter, D.J., Thomas, A., Best, N.G., Lunn, D., 2003. WinBUGS version 1.4.1 User Manual, MRC Biostatistics Unit, Cambridge, UK.
Unger, R., Eder, C., Mayr, J.M., Wernig, J., 2002. Child pedestrian injuries at tram and bus stops. Injury 33 (6), 485–488.
Washington, S.P., Karlaftis, M.G., Mannering, F., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman & Hall/CRC, New York.
World Health Organization (WHO), 2009. Global Status Report on Road Safety: Time for Action. Department of Violence & Injury Prevention & Disability, Geneva, Switzerland.
Wong, S.C., Sze, N.N., Li, Y.C., 2007. Contributory factors to traffic crashes at signalized intersections in Hong Kong. Accident Analysis and Prevention 39, 1107–1113.
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, 443–452.