Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks

Qiang Zeng[a, b], Helai Huang[b, *], Xin Pei[c], S.C. Wong[d]

[a] School of Civil Engineering and Transportation, South China University of Technology, Guangzhou, Guangdong, 510641 P.R. China

[b] Urban Transport Research Center, School of Traffic and Transportation Engineering, Central South University, Changsha, Hunan, 410075 P.R. China

[c] Department of Automation, Tsinghua University, Beijing, P.R. China

[d] Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong

ABSTRACT

This study develops neural network models to explore the nonlinear relationship between crash frequency by severity and risk factors. To eliminate the possibility of over-fitting and to deal with black-box characteristic, a network structure optimization and a rule extraction method are proposed. A case study compares the performance of the modified neural network models with that of the traditional multivariate Poisson-lognormal model for predicting crash frequency by severity on road segments in Hong Kong. The results indicate that the trained and optimized neural networks have better fitting and predictive performance than the multivariate Poisson-lognormal model. Moreover, the smaller differences between training and testing errors in the optimized neural networks with pruned input and hidden nodes demonstrate the ability of the structure optimization algorithm to identify insignificant factors and to improve the model's generalizability. Furthermore, two rule-sets are extracted from the optimized neural networks to explicitly reveal the exact effect of each significant explanatory variable on the crash frequency by severity under different conditions. The rules imply that there is a nonlinear relationship between risk factors and crash frequencies with each injury-severity outcome. With the structure optimization algorithm and rule extraction method, the modified neural network models have great potential for modeling crash frequency by severity, and should be considered a good alternative for road safety analysis.

Keywords: crash frequency by severity; neural network; over-fitting; structure optimization; rule extraction.

* Corresponding author

E-mail address: 641459622@qq.com (Q. Zeng), huanghelai@csu.edu.cn (H. Huang), peixin@mail.tsinghua.edu.cn (X. Pei), hhecwsc@hku.hk (S.C. Wong)

1. Introduction

In the past decade, there has been a quantity of research on predicting crash frequency by certain categories, such as injury severity (e.g., property damage only, possible injury, non-incapacitating injury, incapacitating injury or fatality) (Park and Lord, 2007), the number of vehicles involved (e.g., single vehicle, two vehicles, or three or more vehicles) (Venkataraman et al., 2013) or collision type (e.g., angle, head-on, rear-end, sideswipe or pedestrian-involved) (Ye et al., 2009). The first kind of classification covers most concerns, because crash injury severity is an important aspect in assessing safety performance, in addition to the crash frequency (AASHTO, 2010). Compared with conventional crash prediction models (referred to as "safety performance functions"), modeling crash frequency by severity identifies the effects of observed risk factors (such as the traffic, geometrical and environmental characteristics of sites) on the frequency of accidents with a particular injury-severity outcome. The expected crash frequencies at each level of severity provide deeper insights on the safety situation of a certain road entity (road segment, intersection, etc.). Therefore, while crash totals may not reveal a site deficiency, over exposure of a specific crash severity may uncover otherwise undetected deficiencies. Moreover, the models have been employed to rank road sites with promise for safety improvement, a critical step of network screening in the roadway safety management process (AASHTO, 2010), as injury severity and its associated costs are primary concerns in many programs (Miaou and Song, 2005).

Methodologically, there are mainly two groups of approaches to crash frequency by severity prediction: joint and separate modeling. In the former group, correlation between crash frequencies at various severity levels is the most important issue. To deal with it, a series of techniques have been investigated, such as multivariate regression models (Aguero-Valverde and Jovanis, 2009; Anastasopoulos et al., 2012; Barua et al., 2014, 2016; Bijleveld, 2005; El-Basyouny and Sayed, 2009; El-Basyouny et al., 2014; Ma and Kockelman, 2006; Ma et al., 2008; Park and Lord, 2007), simultaneous equations (Ye et al., 2009, 2013), a joint-probability approach (Pei et al., 2011), two-stage bivariate/multivariate models (Wang et al., 2011; Xu et al., 2014) and multinomial-generalized Poisson models (Chiou and Fu, 2013, 2015; Chiou et al., 2014). The multivariate Poisson regression proposed by Ma and Kockelman (2006) adds a common error term into the Poisson distributions of univariate regressions to account for their correlation, but it does not allow for the commonly observed over-dispersion, and it assumes the identical and positive covariances across crash frequencies (Park and Lord, 2007). In order to improve it, a multivariate Poisson-lognormal regression has been developed (Ma et al., 2008), which is able to accommodate over-dispersion and provides a fully general covariance structure. To account for the spatial correlation among neighboring sites, error terms with Gaussian conditional auto-regressive distribution have been introduced into the multivariate

Poisson-lognormal model (Barua et al., 2014). Based on it, Barua et al. (2016) have proposed a multivariate random parameters count model to further capture unobserved heterogeneity across observations.

Compared with multivariate regression models, the formulation of simultaneous equations, the joint probability model and the two-stage bivariate/multivariate models are less complicated (Pei et al., 2011; Wang et al., 2011; Xu et al., 2014; Ye et al., 2009, 2013). Besides, the computation burden of simultaneous equations is lighter, because their coefficients are calibrated by a simulated likelihood estimation method (Ye et al., 2009), while the others are calibrated by Markov chain Monte Carlo simulation, a typical Bayesian inference method. On the contrary, the multinomial-generalized Poisson models (Chiou and Fu, 2013; Chiou et al., 2014), especially the extension with accommodating spatio-temporal dependence (Chiou and Fu, 2015), are even more complicated than multivariate count models.

Although the correlation across severity levels is significant in many studies, the advantage of joint modeling over separate modeling is not "theoretical" but rather "empirical", as noted by Ma et al. (2008). In the comparative analysis conducted by Lan and Persaud (2012), univariate models are found to fit the crash data better than the multivariate model. Consequently, some researchers continue to separately model crash frequencies at each severity level. For example, Venkataraman et al. (2013) advocate univariate random parameter models to individually predict crash frequency by severity, or other aggregation types, by accounting for heterogeneities across unobserved or unobservable factors. All of the above-mentioned models are based on a generalized linear function framework and certain assumed distributions of crash data. However, in some cases, these assumptions may be violated and thereby result in biased inferences (Li et al., 2008).

Relative to the statistical models, without any prior knowledge or assumption on model structure, some artificial intelligence models can be used to approximate the underlying nonlinear relationship between crash frequency by severity and safety predictors (Haykin, 2009). As a common class of artificial intelligence models, neural network models have been successfully used in many fields of transportation research (Karlaftis and Vlahogianni, 2011). For highway safety analysis, a number of studies have investigated the performance of neural network models in predicting crash frequency or injury severity (Abdelwahab and Abdel-Aty, 2001; Chang, 2005; Huang et al., 2016; Zeng and Huang, 2014b). The results show that neural network models outperform some traditional statistical models, such as the negative binomial model of crash frequency prediction and the ordered logit/probit models of crash injury severity prediction. To the best of our knowledge, neural networks have not yet been employed to predict crash frequency by severity.

Moreover, with the development of neural network techniques, the commonly criticized weaknesses of crash prediction, the over-fitting problem and the black-box characteristic, have been mostly eliminated. Advanced methods for network training and structure optimization can establish generalized neural network models that

effectively approximate the relationship between crash frequency by severity and explanatory variables (Haykin, 2009). In addition, piecewise linear rules extracted from the developed neural networks are able to clearly illustrate the effects of risk factors (Setiono and Thong, 2004).

In summary, this study attempts to develop advanced neural networks for modeling the nonlinear relationship between crash frequency by severity and risk factors, and to clarify the effects of factors on the outcomes by extracting rules from the developed neural networks. To demonstrate the proposed methods, the neural network models are compared with the multivariate Poisson-lognormal model with regard to fitting and predictive performance. Accordingly, the remainder of this paper is organized as follows. The next section specifies the proposed models and methods. The collected data for model demonstration are described in Section 3. Section 4 introduces the detailed implementation of the proposed models and discusses the results. Finally, conclusions and recommendations for future research are presented in Section 5.

2. Methodology

The multivariate Poisson-lognormal model, one of the most widely used statistical models for jointly predicting crash frequency and severity, is used as a benchmark in this study to compare its fitting and predictive performance with those of the proposed neural network models. In this section, the model architectures of the multivariate Poisson-lognormal and neural network models are specified. Then, the training, structure optimization, and rule extraction algorithms for the neural network models are described.

2.1. Model specification

2.1.1. Multivariate Poisson-lognormal model

In the multivariate Poisson-lognormal model, the crash count Y_{its} at site i during period t at injury severity degree s is assumed to follow a Poisson distribution (Ma et al., 2008), given λ_{its}, that is,

$$\mathrm{P}(Y_{its} = y_{its} \,|\, \lambda_{its}) = \lambda_{its}^{y_{its}} \frac{e^{-\lambda_{its}}}{y_{its}!},$$

$$i = 1,2,\cdots N,\ \ t = 1,2,\cdots T,\ \ s = 1,2,\cdots S,\ \ y_{its} = 0,1,2,3,\cdots, \qquad (1)$$

where N, T, and S are the number of observed sites, the periods and the categorized injury severity levels, respectively. The mean of Y_{its}, λ_{its}, is assumed to

have a generalized linear relationship with the explanatory variables, $\mathbf{X}_{it}$, such that

$$\ln \lambda_{its} = \mathbf{X}_{it}'\boldsymbol{\beta}_s + \varepsilon_{its}, \tag{2}$$

in which $\boldsymbol{\beta}_s$ are the coefficients to be estimated. The error term ε_{its} accommodates the crash severity correlation and the common over-dispersion, which is multi-normally distributed as

$$\boldsymbol{\varepsilon}_{it} \sim N_S(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\varepsilon}_{it} = \begin{pmatrix} \varepsilon_{it1} \\ \varepsilon_{it2} \\ \cdots \\ \varepsilon_{itS} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1S} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2S} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{S1} & \sigma_{S2} & \cdots & \sigma_{SS} \end{pmatrix}. \tag{3}$$

2.1.2. Neural network model

Neural network models are information-processing mechanisms that are inspired by biological nervous systems (Haykin, 2009). Various categories of neural network models have been developed with different network architectures, such as the radial basis function and the self-organizing feature map. The multilayer perceptron, which is known as a universal approximator and is the most popular neural network for data mining, is used to model the underlying nonlinear relationship between crash frequency by severity and risk factors in this study. Although they can be predicted simultaneously in the same neural network, the crash frequencies at each injury severity are modeled separately herein to identify their respective pertinent predictors. Fig. 1 shows the structure of the developed multilayer perceptrons with fully connected neurons.

Consider a dataset containing N_1 continuous attributes and N_2 categorical attributes that may affect the crash frequency at the severity level $s(s = 1, 2, \cdots, S)$. As in many statistical modeling methods, each categorical attribute $A_n(n = 1, 2, \cdots, N_2)$ is transformed into $m_n - 1$ binary attribute(s) $a_1^n, \cdots a_j^n, \cdots, a_{m_n-1}^n$, where m_n is the number of possible values for A_n. $a_j^n = 1$ if A_n is equal to category j; and $a_j^n = 0$ otherwise. Each of the transformed attributes, together with the continuous attributes, is represented by a node $x_i(i = 2, \cdots I)$ in the input layer. In addition, an input node, with $x_1 = 1$, is added. The weights of its connections with hidden neurons

are the biases. Therefore, the number of units I in the input layer is

$$I = N_1 + \sum_{n=1}^{N_2} (m_n - 1) + 1 . \quad (4)$$

To fit the training data well, the number of neurons in the hidden layer must be sufficiently large. If it is assumed to be J, then the connection weight between hidden node $j(j = 1, \cdots, J)$ and input node $i(i = 1, \cdots, I)$ is $w_{j,i}^{(1)}$. The hyperbolic function, $\tanh(\cdot)$, which is an odd sigmoid transfer function, is used for all of the hidden nodes. In the output layer, the only unit, ψ, represents the expected crash frequency at severity level s. $w_j^{(2)}$ denotes the weight of the connection between the output node and the hidden node $j(j = 1, \cdots, J)$. A linear function is employed as the transfer function for the output node. Then, the expected crash frequency at severity level s is given by,

$$\psi = \sum_{j=1}^{J} w_j^{(2)} \tanh\left(\sum_{i=1}^{I} w_{j,i}^{(1)} x_i\right) . \quad (5)$$

2.2. Network training

The conjugate gradient algorithm, which has a better learning performance than the popular back-propagation algorithm (Haykin, 2009), is adopted in this study to train the neural networks. For the collected samples $\{\mathbf{x}(m), o(m) \mid m = 1, 2, \cdots, M\}$ of severity $s(s = 1, 2, \cdots, S)$, where $\mathbf{x}(m)$ and $o(m)$ are a vector of risk factors and the corresponding observed crash frequency at severity s, respectively, and M is the number of samples, the conjugate gradient updates the connection weight vector $\mathbf{w}$ as follows:

$$\begin{aligned} \mathbf{w} &= (w_1, \cdots, w_{(j-1)I+i}, \cdots w_{JI}, w_{JI+1}, \cdots, w_{JI+j}, \cdots w_{J(1+I)}) \\ &= (w_{1,1}^{(1)}, \cdots, w_{j,i}^{(1)}, \cdots w_{J,I}^{(1)}, w_1^{(2)}, \cdots, w_j^{(2)}, \cdots w_J^{(2)}) \end{aligned} .$$

1. Randomly select $w_{j,i}^{(1)}(j = 2, \cdots, J;\ i = 1, \cdots, I)$ and $w_j^{(2)}(j = 1, \cdots, J)$ from two uniform distributions. The means of both distributions are equal to 0, and their variances are $1/J$ and 1, respectively. Set the initial iteration $t = 0$.

2. According to weight vector $\mathbf{w}(0)$, calculate the expected network outputs, $\psi_m(m = 1, 2, \cdots, M)$, the derivative of outputs on all weights,

$\frac{\partial \psi(m)}{\partial \mathbf{w}(0)} (m = 1, 2, \cdots, M)$, and the gradient vector, $\mathbf{g}(0)$:

$$\frac{\partial \psi(m)}{\partial w_h} = \begin{cases} \frac{\partial \psi(m)}{\partial w_j^{(2)}} = \tanh(\sum_{i=1}^{I} w_{j,i}^{(1)} x_i(m)), & if \ w_h = w_j^{(2)} \\ \frac{\partial \psi_k(m)}{\partial w_{j,i}^{(1)}} = w_j^2 \tanh'(\sum_{i=1}^{I} w_{j,i}^{(1)} x_i(m)) x_i(m), & f \ w_h = w_{j,i}^{(1)} \end{cases}, \tag{6}$$

$$\mathbf{g}(t) = \frac{1}{N} \sum_{m=1}^{M} [o(m) - \psi(m)] \frac{\partial \psi(m)}{\partial \mathbf{w}(t)}. \tag{7}$$

3. Set $\mathbf{s}(0) = \mathbf{r}(0) = -\mathbf{g}(0)$.

4. In iteration t, for the fixed $\mathbf{w}(t)$ and $\mathbf{s}(t)$, use the advance-and-retreat method to linearly search the optimal $\eta(t)$ by minimizing the cost function, $\xi_{av}(\mathbf{w}(t) + \eta \mathbf{s}(t))$:

$$\xi_{av}(\mathbf{w}) = \frac{1}{2M} \sum_{m=1}^{M} [o(m) - \psi(m)]^2. \tag{8}$$

5. Check the convergence criteria. If the Euclidean norm of $\mathbf{r}(t)$ decreases to a certain small portion, ε, of its initial value, $\|\mathbf{r}(0)\|$, or the iteration number meets its maximum value, T, the algorithm is done:

$$\|\mathbf{r}(n)\| \le \varepsilon \|\mathbf{r}(0)\|, \text{ or } t = T.$$

6. Update the connection weight vector:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \eta(t) \mathbf{s}(t). \tag{9}$$

7. Calculate the gradient vector $\mathbf{g}(t+1)$ by formulas (6)-(7) according to $\mathbf{w}(t+1)$. Set $\mathbf{r}(t+1) = -\mathbf{g}(t+1)$.

8. Calculate $\beta(t+1)$ by the Polak-Ribiere method:

$$\beta(t+1) = \max\{\frac{\mathbf{r}'(t+1)(\mathbf{r}(t+1) - \mathbf{r}(t))}{\mathbf{r}'(t)\mathbf{r}(t)}, 0\}. \tag{10}$$

9. Update the direction vector:

$$\mathbf{s}(t+1) = \mathbf{r}(t+1) + \beta(t+1)\mathbf{s}(t). \tag{11}$$

10. Set $t = t+1$, and return to step 4.

2.3. Structure optimization

Following Setiono and Leow (2000), the structure optimization algorithm, which has been successfully used to develop an optimized neural network model for crash injury severity prediction (Zeng and Huang, 2014b), is proposed to improve the generalization capacity of the neural network models and to identify the insignificant explanatory variables. This method prunes the nodes that do not cause any significant deterioration of the networks' accuracy. The mean absolute deviations of the training set $\mathbf{T}$ and testing set $\mathbf{X}$, that is, p and q, are used to evaluate the fitting and predictive performance during network optimization:

$$p = \frac{1}{M_1} \sum_{o(m)\in\mathbf{T}} \left| o(m) - \psi(m) \right|, \tag{12}$$

$$q = \frac{1}{M_2} \sum_{o(m)\in\mathbf{X}} \left| o(m) - \psi(m) \right|, \tag{13}$$

where M_1 and M_2 are the number of samples in the training and testing sets, respectively.

The following steps describe the detailed pruning process.

1. Train the network with a relatively large number of hidden nodes using the conjugate gradient algorithm.
2. Calculate the p and q of the trained neural network, and set $p\_b = p$, $q\_b = q$, and $ermax = \max\{p\_b, q\_b\}$.
3. For each $i (i = 1, \cdots, I)$, set $w_{i,j}^{(1)} = 0 (j = 1, \cdots, J)$ and calculate the fitting errors p_i.
4. Retrain the network with $w_{l,j}^{(1)} = 0 (j = 1, \cdots, J)$, where $p_l = \min_i p_i$, and compute p and q for the retrained network.
5. If $p \le (1+\sigma)ermax$ and $q \le (1+\sigma)ermax$, then remove the input node l, set $p\_b = \min\{p, p\_b\}$, $q\_b = \min\{q, q\_b\}$, $ermax = \max\{p\_b, q\_b\}$, $I = I - 1$, and go back to step 3; otherwise, keep the previous weights of the network connections.
6. For each $j (j = 1, \cdots, J)$, set $w_j^{(2)} = 0$ and calculate the fitting errors p_j.

7. Retrain the network with $w_h^{(2)} = 0$, where $p_h = \min_j p_j$, and compute p and q of the retrained network.

8. If $p \le (1+\sigma)ermax$ and $q \le (1+\sigma)ermax$, then remove the hidden node h. Set $p\_b = \min\{p, p\_b\}$, $q\_b = \min\{q, q\_b\}$, $ermax = \max\{p\_b, q\_b\}$, and $J = J - 1$, and go back to step 6; otherwise, keep the previous weights of the network connections.

In the above process, p_b and q_b represent, respectively, the minimal mean absolute deviations of the training and testing sets achieved so far. During the pruning process, generally, p_b increases while q_b decreases. $ermax$ is used to determine whether or not a node can be removed to remove as many insignificant nodes as possible without sacrificing the generalization accuracy. In addition, σ is the margin by which the error is allowed to increase when pruning a certain node.

2.4. Rule extraction

The rule extraction method developed by Setiono and Thong (2004) is modified to generate exact and comprehensible rules from the pruned neural network to illustrate the effects of significant explanatory variables. In the next subsections, a particle swarm optimization algorithm-based approach to approximating the transfer functions of hidden units is introduced as a critical step in the method, and then the rule extraction process is described.

2.4.1. Approximating transfer functions

The transfer functions of the hidden nodes can be approximated by piecewise functions. Theoretically, the more pieces fit the function, the more accurate the rule set, and the more rules may be extracted. To balance the two aspects, a three-piece linear function suggested by Setiono and Thong (2004) is used to approximate the transfer function of each hidden node $j (j = 1, \cdots, J)$, $\tanh(\cdot)$, as shown in Fig. 2. The slopes, β_{j0} and β_{j1}, and the cut-off point, ξ_{j0}, are three undetermined parameters that minimize the sum of the squared deviations,

$$\min \sum_{m=1}^{M} (\tanh(\upsilon_j(m)) - L_j(\upsilon_j(m)))^2 , \tag{14}$$

where

$$L_j(x) = \begin{cases} -\alpha_{j1} + \beta_{j1} x & if\ x < -\xi_{j0} \\ \beta_{j0} x & if\ -\xi_{j0} \le x \le \xi_{j0} \\ \alpha_{j1} + \beta_{j1} x & if\ x > \xi_{j0} \end{cases}, \tag{15}$$

$$\upsilon_j(m) = \sum_{i=1}^{I} w_{j,i}^{(1)} x_i(m), \tag{16}$$

$$\alpha_{j1} = (\beta_{j0} - \beta_{j1})\xi_{j0}. \tag{17}$$

2.4.2. Searching the optimal parameters

To approximate the transfer function accurately, the particle swarm optimization algorithm, an efficient global search method, is used to solve the preceding nonlinear optimization problem. The particle swarm optimization algorithm is well-known for its exploration capacity, its exploitation capacity and its easy implementation (Poli et al., 2012). In the algorithm, each feasible solution $(\beta_{j0}, \beta_{j1}, \xi_{j0})$ is referred to as a "particle", $\mathbf{U}$, and each particle flies around the three-dimensional search space with a velocity $\mathbf{V}$, which is updated iteratively according to the best solution of the particle achieved so far (particle best, **pbest**) and the best solution obtained by all of the particles in the swarm so far (global best, **gbest**):

$$\mathbf{V}_s^{r+1} = \mathbf{V}_s^r + c_1\lambda_1(\mathbf{pbest}_s^r - \mathbf{U}_s^r) + c_2\lambda_2(\mathbf{gbest}_s^r - \mathbf{U}_s^r), \tag{18}$$

$$\mathbf{U}_s^{r+1} = \mathbf{U}_s^r + \mathbf{V}_s^{r+1}, \tag{19}$$

$$\mathbf{U} = (\beta_{j0}, \beta_{j1}, \xi_{j0});\ r = 1, 2, \cdots, R;\ s = 1, 2, \cdots, S.$$

where $\mathbf{U}_s^r$ is the s th particle at the r th iteration, and $\mathbf{V}_s^{r+1}$ is its flying velocity to the $r+1$ th iteration. c_1 and c_2 are two acceleration constants, while λ_1 and λ_2 are two uniform random numbers in $[0,1]$. R is the maximum iteration number, and S is the number of particles used for searching the optimal solution.

2.4.3. Generating regression rules

Once the transfer functions of the hidden units have been approximated, the relationship between the network inputs and outputs can be formulated with piecewise linear functions. The detailed steps for extracting rules from the optimized neural network are as follows:

1. For each hidden unit $j(j = 1, \cdots, J)$, generate a three-piece linear function $L_j(x)$ with the approach previously described.
2. According to the pair cut-off points in $L_j(x)$, $-\xi_{j0}$ and ξ_{j0}, a certain input can be located in one of three sections of hidden node j. Then, J hidden nodes will

result in $\underbrace{3\times3\times\cdots\times3}_{J}$ locations for inputs. Consequently, the whole input space can be separated into 3^J subspaces.

3. For each non-empty subspace, the rule consequence is $\tilde{y}=\sum_{j=1}^{J} w_j^{(2)}\cdot L_j(\upsilon_j)$, where $\upsilon_j=\sum_{i=1}^{I} w_{i,j}^{(1)}\cdot x_i$, and the rule condition is $C_1 \& C_2 \& \cdots C_J$, where C_j is either $\upsilon_j<-\xi_{j0}$, $-\xi_{j0}\le\upsilon_j\le\xi_{j0}$ or $\upsilon_j>\xi_{j0}$.

3. Data preparation and preliminary analysis

A crash dataset obtained from the Traffic Information System maintained by the Transport Department of Hong Kong is used to demonstrate the proposed neural network models and to compare them with the multivariate Poisson-lognormal model. This dataset contains 211 road segments that are evenly and widely distributed across Hong Kong. Geographical information system techniques are used to map crashes to these segments. The injury severity outcomes of the crashes are divided into two levels, fatality or serious injury and slight injury, and the annual crash numbers at each severity level at each site during 2002 to 2006 are obtained. The road geometric and traffic information is also included in the dataset. Table 1 illustrates the definitions and descriptive statistics of the variables used in the model development.

The lane changing opportunity (LCO) variable refers to the different types of central lane marking, with values 0, 1 and 2 representing, respectively, double continuous lines, double lines with one continuous line and one dashed line, and a single dashed line. For those sub-segments with more than one type of central lane marking, the length-weighted average values are used. Pei et al. (2012) provides a more detailed description of the lane changing opportunity.

According to Table 1, the mean and variance of crash frequency at the slight injury level are 6.04 and 25.81, respectively, indicating a possible over-dispersion. A similar characteristic is found in the fatality or serious injury crash frequency. In the multivariate Poisson-lognormal model, to account for the potential nonlinear relationship between crash frequencies and traffic volumes, the natural logarithm of AADT and Length, ln(AADT) and ln(Length), are modeled as other factors (Zeng and Huang, 2014a).

Correlation tests and multi-collinearity diagnoses for the risk factors are then conducted. According to the results of the Pearson correlation tests, we find that ln(AADT) and Lane, ln(AADT) and Park, Lane and LCO, SL and Shoulder, SL and Park are significantly correlated with correlation coefficients greater than 0.6. To reduce the model complexity, Lane, Park, and Shoulder are therefore excluded from

the models. The results of the diagnoses indicate that there is no significant collinearity in the remaining factors.

4. Model implementation and result analysis

4.1. Model implementation

The multivariate Poisson-lognormal model is estimated with the freeware WinBUGS, which is a popular platform to make Bayesian inference and is well-known for its flexible programming environment (Zeng and Huang, 2014a). In the absence of sufficient prior knowledge, non-informative priors are specified for the parameters and the hyper-parameters. Specifically, a diffused normal distribution $N(0,10^4)$ is used as the priors of all elements of $\boldsymbol{\beta}_s\,(s=1,2)$, while a Wishart prior $W(\mathbf{P},r)$ is used for $\boldsymbol{\Sigma}^{-1}$, where $\mathbf{P}=\begin{bmatrix}1, & 0\\ 0, & 1\end{bmatrix}$ represents the scale matrix and $r=2$ is the degrees of freedom (El-Basyouny and Sayed, 2009; Park and Lord, 2007). Five hundred thousand iterations of the Markov chain Monte Carlo simulation are made, with the first 4000 iterations acting as burn-ins. After ensuring the Markov chain Monte Carlo convergence by the Gelman-Rubin statistics available in WinBUGS, another 50,000 iterations are set to make summaries for the (hyper-) parameters.

The training, the structure optimization, and the rule extraction algorithms of the neural network models are programmed in MATLAB. All of the variables are normalized for the convenience of network training. To compare the performance of the models fully, a 5-fold cross validation is conducted, where the dataset is randomly divided into five parts with equal number of observations/patterns. Each time, the sub-dataset of any four parts is input for training the models while the rest is used for testing the predictive performance. Based on the collected data, $I=14$, we first set $J=10$ for all networks. In the network training, $\varepsilon=0.001$ and $T=50$. We assume that $\sigma=0.05$ in the structure optimization algorithm, while $R=300$ and $S=700$ in the particle swarm optimization algorithm.

4.2. Model comparison

The results of the model comparison are summarized in Table 2. With regard to the five folds of model comparison, in terms of the mean absolute deviation criteria, all of the trained and optimized neural network models have lower fitting and predictive errors for the training and testing datasets than the multivariate Poisson-lognormal models, at both the fatality or serious injury and the slight injury levels. This demonstrates that neural network models of crash frequency prediction may give a better approximation performance than certain traditional statistical

models, which is probably due to the neural network's capacity for approximating arbitrary nonlinear functions.

After pruning the network structure with the structure optimization algorithm, the model fitting is generally expected to be degraded to some extent but the model prediction should be improved as discussed in the section 2.3. But in the results as shown in Table 2, it is surprisingly found that both the fitting and predictive errors of the neural network models are reduced by the proposed model structure optimization algorithm. As generally known, like other training algorithms, the proposed conjugate gradient algorithm may sometimes be locally converged (Haykin, 2009). Therefore, a presumable cause for the reduced model-fitting errors may be that pruning nodes and retraining network could help to escape from local minima and to search for better solutions. As a result, we may argue that the model generalization performance associated with the proposed algorithm is improved as reflected by the reduced model fitting and predictive errors.

Moreover, certain numbers of input and hidden nodes are removed from the trained neural networks in all of the five folds, which indicates that the original models have redundant nodes, and that the factors corresponding to those removed input nodes may have no significant effects on the crash frequency by severity.

It is also noticeable that the five pairs of optimized neural networks end up with slight distinctions in their mean absolute deviation values and the final sets of input and hidden nodes. This instability is presumably attributable to the small sample size (Xie et al., 2007), given the important impact of sample size on a model's generalizability (Haykin, 2009).

4.3. Interpretation of the explanatory variables

The specific conditions and consequences of the rules extracted from the optimized neural networks are shown in Tables 3-6. In these tables, we can clearly see the effects of the significant factors on crash frequencies at the two levels of injury severity, under diverse conditions. For the purpose of comparison, the estimation results of the parameters and the hyper-parameters in the multivariate Poisson-lognormal model are shown in Table 7 and Table 8, respectively. According to the results in Table 8, we see that both the fatality or serious injury and the slight injury crash data are over-dispersed, as their extra-Poisson variations (σ_{11} and σ_{22}) are significantly positive at the 95% credible level. Moreover, the correlation coefficient $\rho\,(=\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}})$ reaches 0.763, showing that the crash frequencies at the two injury levels are highly correlated.

In this section, we analyze mainly the rule consequences in Tables 5 and 6, as the rule conditions in Tables 3 and 4 may be difficult to understand. Instead, we employ the characteristics of the road segments involved at certain particular rules to illustrate

the effects of the risk factors. Even so, it is noticeable that, based on the conditions, the rule to which each observation in the analysis should be assigned can be determined accurately. Comparing the results in Tables 5 and 6 with those in Table 7, we find that the coefficients of all of the identified factors in the optimized neural networks are significant at the 95% credible level in the multivariate Poisson-lognormal model, except Rainfall in the fatality or serious injury neural network.

Regarding the main effects of the risk factors identified, most of the risk factors have consistent signs, as shown in Tables 5 and 6, which also conform to the signs in the multivariate Poisson-lognormal model results shown in Table 7. The signs of the coefficients of the factors AADT, Length, SL, BS and Rainfall in the slight injury neural network, and AADT and Diverge in the fatality or serious injury neural network are identical at all rules. As for the other factors, it is interesting to find a few different signs in several specific rules. Moreover, it is observed that the estimated coefficient values are also distinct for several specific rules. This implies that those risk factors probably have variable safety effects under different road conditions. This could be important evidence of nonlinear relationship between crash frequency by severity and the risk factors, which cannot be identified and modeled with the traditional generalized linear regression models, such as the multivariate Poisson-lognormal model.

According to the results in Table 5, more slight injury crashes tend to occur on longer road segments with more daily traffic, as observed by the positive coefficient estimations associated with all of the eleven rules for AADT and Length. This is a reasonable conclusion, given that AADT and segment length are always used as two of the crash exposure variables in highway safety analysis (AASHTO, 2010; Zeng and Huang, 2014a). Nonetheless, the proposed neural network model presents specific values for varied safety effects under different conditions. For example, increasing one unit of AADT is expected to increase only 0.08 crashes (based on the normalized data) under Condition 10, but 2.11 crashes (almost 26 times the former) under Conditions 1 and 2.

Slight injury crash frequencies are found lower on road segments with higher speed limits. It may be attributed to two reasons: (1) roadway segments designed for higher speeds are usually well planned, constructed, and managed, features that promote road safety, as argued by some previous researchers (Milton and Mannering, 1998), and (2) given a collision occurs, higher speed usually increases the likelihood of severe injury and fatality while decreases that of slight injury (Zeng and Huang, 2014b).

The presence of median barriers is found to reduce slight injury crash occurrence at most rules. A number of existing studies have also found that median barriers can effectively prevent cross-median crashes (Donnell and Mason, 2006). However, the estimated coefficients are positive at Rules 6 and 9. For those observations at these rules, about 90 % of the road segments have median barriers, of which most are

inner-city highways with heavy daily traffic (mean = 36,538 vehicles) and many merging ramps (mean = 1.74). These factors may hinder safe driving and bring about more slight injury collisions related to median barriers. Under all conditions, the presence of bus stops decreases slight injury crash frequencies, which may be attributed to the increased interaction between buses and other vehicles when entering or leaving bus bays (Pei et al., 2012).

The negative coefficients of Gradient under most conditions indicate that more slight crashes are expected to occur on road segments with steeper downgrade slopes, which is generally consistent with engineering experience. Besides, Gradient is found to decrease the crash frequencies at Rules 6 and 10. Most of the involved road segments are very long, such as Tsing Long Highway (9.07 km), Shek O Road (7.75 km) and Tolo Highway (5.60 km). Driving on the downgrade directions of these long highways, drivers may be more careful, thus reducing the crash risk.

Slight injury crash frequencies usually increase with more lane changing opportunities. Lane-cutting maneuvers often increase vehicle interaction, such as overtaking, thereby raising the incidence of traffic conflict (Pei et al., 2012). It is interesting to find that more lane changing opportunities could bring about more slight injury collisions under Condition 10. The referred roadways mainly consists of freeways, such as the longest segment in the dataset— Tsing Long Highway (9.07 km). Lane changing maneuver is less frequent on these freeways than on those busy inner-city roadways, which may reduce the vehicle speed variance. This may possibly explain why LCO negatively affects the slight injury crash frequency on them.

Generally, rainfall impairs visibility and makes road surfaces slippery, thereby reducing skidding resistance, which raises the probability of crash occurrence. This is why Rainfall has positive model coefficients in Table 5, which indicates that rainfall may lead to more slight injury crashes (Pei et al., 2012).

Based on the results in Table 6, we find that more fatality or serious injury crashes are associated with longer roadway segments, more daily traffic, no median barrier, presence of bus stop, steeper downgrades and more precipitation under most or all conditions, which is similar to the results of slight injury crashes. The negative coefficients for the variable Diverge indicate that more diverging ramps give rise to a higher fatality or serious injury crash risk, which may be attributable to more conflicts at the sites approaching diverging ramps.

At Rule 21, the length is found negatively related to fatality or serious injury crash frequency, in which all fatality or serious injury crashes occurred on Tsing Long Highway. For the longest road segment, its average annual fatality or serious injury crash number is only 1.3, smaller than the mean of the whole population (1.8). A possible reason for the negative coefficient of Length may be that some unobserved factors (such as well design and maintenance) associated with this highway greatly promote the safety situation. Under the same Condition, Gradient is found positively related to the fatality or serious injury crash frequency. It may be a result of a similar reason to the corresponding findings in slight injury crashes, that is, drivers are

usually more cautious when driving on the downgrade of the so long (9.07 km) segment.

Regarding the observations at Rules 16 and 17, most of the road segments are also inner-city highways with heavy daily traffic (mean = 35,462 vehicles) and many merging ramps (mean = 1.82). Like the situation at Rules 6 and 9, these factors may impede safe driving and result in more fatality or serious injury median-related crashes. Meanwhile, there are bus stops on all of the roadway segments at Rules 16 and 17. The decreased travel speed of buses entering or leaving bus bays could reduce the probability of severe crashes. As a consequence, the presence of bus stops decreases the fatality or serious injury crash frequency on the segments.

Probably due to the same reason as for the slight injury crashes, Rainfall has positive coefficients under most conditions. However, drivers tend to be more careful and reduce their speed when driving on rainy areas. That may be why Rainfall is negatively related to the fatality or serious injury crash frequency under Conditions 14 and 15, since the annual precipitation of most involved observations are over 3000 mm.

5. Conclusions and future research

This study develops advanced neural networks for modeling the nonlinear relationship between crash frequency by severity and the related factors. To improve the generalization capacity and to handle the black-box characteristic of neural networks, a structure optimization algorithm and a modified rule extraction algorithm are proposed. A crash dataset obtained from the Traffic Information System maintained by the Transport Department of Hong Kong, where crashes are classified into slight injury and fatality or serious injury severity degrees, is used to demonstrate the proposed methods and to compare them with the results of a multivariate Poisson-lognormal model.

Despite the over-dispersed crash data and the high correlation between the crash frequencies of the different injury degrees, the results show that both the trained and the optimized neural networks outperform the multivariate Poisson-lognormal model in fitting and predictive performance. It indicates the neural network's superiority over the multivariate Poisson-lognormal model in modeling crash frequency by severity. When several input and hidden nodes are deleted from the original neural networks, better approximation performance is achieved, demonstrating the structure optimization algorithm's ability to identify insignificant factors and to improve the model's generalization capacity. The optimized neural networks generate two rule-sets in which the coefficients of the explanatory variables are different, which confirms that they are nonlinearly related to the crash frequencies. The signs of these coefficients have identical directions under most conditions, and are consistent with those in the multivariate Poisson-lognormal model. Moreover, most of the results for the explanatory variables are reasonable and conform to traffic engineering

experience or the findings of previous studies, which further validates the proposed methods.

It is worth noting that the other aforementioned statistical model may have better performance than the multivariate Poisson-lognormal model for the collected data in this study, although the latter is the most popular method for jointly modeling crash frequency and severity. For example, the identified nonlinear relationship between crash frequency by severity and risk factors could be viewed as unobserved heterogeneities across observations. The heterogeneities could be accommodated in a multivariate random parameters Poisson-lognormal model, and the empirical analysis based on our collected dataset indicates that it is potentially a better fitting approach. Further research efforts could be made to compare the proposed neural network models with the emerging advanced statistical models on more field datasets. Further, as mentioned above, the developed neural network models can be employed as an alternative approach for identifying sites with promise for improving safety. In the absence of the averaged crash cost of each injury degree level, this part of the application has not been conducted. More comprehensive crash data are needed to compare the proposed neural network techniques with the state-of-the-art methods, such as the Bayesian hierarchical models, for site ranking.

Acknowledgements

This research was jointly supported by the Natural Science Foundation of China (No. 71371192, 71301083), the Hong Kong Research Grants Council of the Hong Kong Special Administrative Region, China (No. 717512) and a grant from the Joint Research Scheme of National Natural Science Foundation of China/Research Grants Council of Hong Kong (No. 71561167001 & N_HKU707/15).

References

Highway Safety Manual, 1st edition, 2010. AASHTO, Washington D.C.

Abdelwahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record 1746, 6-13.

Aguero-Valverde, J., Jovanis, P.P., 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. Transportation Research Record 2136, 82-91.

Anastasopoulos, P. C., Shankar, V. N., Haddock, J. E., Mannering, F. L., 2012. A multivariate tobit analysis of highway accident-injury-severity rates. Accident Analysis and Prevention 45, 110-119.

Barua, S., El-Basyouny, K., Islam, M. T., 2014. A full Bayesian multivariate count data model of collision severity with spatial correlation. Analytic Methods in Accident Research 3, 28-43.

Barua, S., El-Basyouny, K., Islam, M. T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Analytic Methods in Accident Research 9, 1-15.

Bijleveld, F. D. 2005. The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accident Analysis and Prevention 37 (4), 591-600.

Chang, L., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Safety Science 43 (8), 541-557.

Chiou, Y. C., Fu, C., 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. Accident Analysis and Prevention 50, 73-82.

Chiou, Y. C., Fu, C., Chih-Wei, H., 2014. Incorporating spatial dependence in simultaneously modeling crash frequency and severity. Analytic methods in accident research 2, 1-11.

Chiou, Y. C., Fu, C., 2015. Modeling crash frequency and severity with spatiotemporal dependence. Analytic Methods in Accident Research 5, 43-58.

Donnell, E.T., Mason Jr, J.M., 2006. Predicting the frequency of median barrier crashes on Pennsylvania interstate highways. Accident Analysis and Prevention 38(3), 590-599.

El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention 41 (4), 820-828.

El-Basyouny, K., Barua, S., Islam, M. T., 2014. Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accident Analysis and Prevention 73, 91-99.

Haykin, S.S., 2009. Neural networks and learning machines, 3rd edition. Prentice Hall, New York.

Huang, H., Zeng, Q., Pei, X., Wong, S.C., Xu, P., 2016. Predicting crash frequency using an optimized radial basis function neural network model. Transportmetrica A 12 (4): 330-345.

Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transportation Research Part C 19 (3), 387-399.

Labi, S., 2011. Efficacies of roadway safety improvements across functional subclasses of rural two-lane highways. Journal of Safety Research 42 (4), 231-239.

Lan, B., Persaud, B., 2012. Evaluation of multivariate Poisson log normal Bayesian methods for before-after road safety evaluations. Journal of Transportation Safety and Security 4 (3), 193-210.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accident Analysis and Prevention 40 (4), 1611-1618.

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 (3), 964-975.
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 (4), 699-720.
Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25 (4), 395-413.
Park, E.S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, 1-6.
Pei, X., Wong, S.C., Sze, N.N., 2011. A joint-probability approach to crash prediction models. Accident Analysis and Prevention 43 (3), 1160-1166.
Pei, X., Wong, S.C., Sze, N.N., 2012. The roles of exposure and speed in road safety analysis. Accident Analysis and Prevention 48, 464-471.
Poli, R., Kennedy, J., Blackwell, T., 2007. Particle swarm optimization. Swarm intelligence 1, 33-57.
Setiono, R., Leow, W.K., 2000. Pruned neural networks for regression. PRICAI 2000 Topics in Artificial Intelligence. Springer Berlin, Heidelberg, pp. 500-509.
Setiono, R., Thong, J.Y., 2004. An approach to generate rules from neural networks for regression problems. European Journal of Operational Research 155, 239-250.
Venkataraman, N., Ulfarsson, G.F., Shankar, V.N., 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. Accident Analysis and Prevention 59, 309-318.
Wang, C., Quddus, M. A., Ison, S. G., 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. Accident Analysis and Prevention 43 (6), 1979-1990.
Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural networks: an empirical analysis. Accident Analysis and Prevention 39 (5), 922-933.
Xu, X., Wong, S.C., Choi, K., 2014. A two-stage bivariate logistic-Tobit model for the safety analysis of signalized intersections. Analytic Methods in Accident Research 3-4, 1-10.
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.
Ye, X., Pendyala, R. M., Shankar, V., Konduri, K. C., 2013. A simultaneous equations model of crash frequency by severity level for freeway sections. Accident Analysis and Prevention 57, 140-149.
Zeng, Q., Huang, H., 2014a. Bayesian spatial joint modeling of traffic crashes on an urban road network. Accident Analysis and Prevention 67, 105-112.

1 Zeng, Q., Huang, H., 2014b. A stable and optimized neural network model for crash
2 injury severity prediction. Accident Analysis and Prevention 73, 351-358.

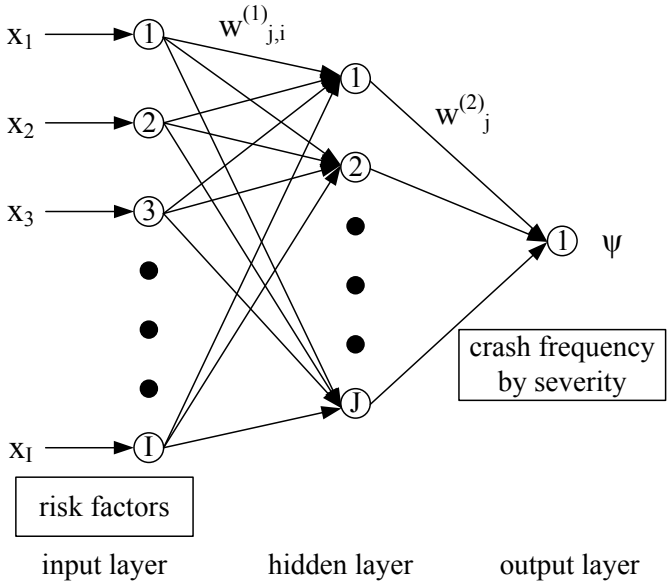

Fig. 1. Developed multilayer perceptron structure

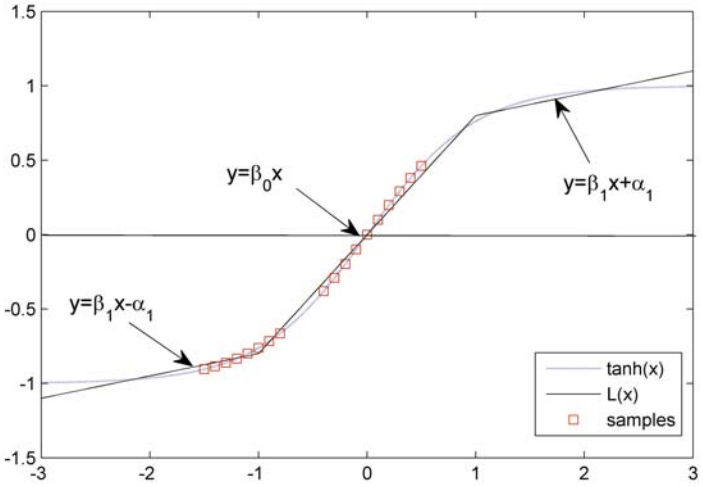

Fig. 2. Three-piece linear approximation of tanh(·)

Table 1 Descriptive statistics of the variables

Variable	Description	Mean	SD	Min.	Max.
Response variable					
Slight	Slightly injured crash count per segment per year	6.04	5.08	0	40
KSI	Killed and seriously injured crash count per segment per year	1.60	1.90	0	12
Explanatory variables					
AADT	Average annual daily traffic (vehicles)	22077	19945	1164	101632
Length	Segment length (km)	1.47	1.55	0.15	9.07
Lane	Number of lanes	2.41	1.18	1	7
Width	Average width of each lane (m)	3.63	0.64	2.40	7.30
SL	Posted speed limit (km/h)	60.3	14.7	50	110
Merge	Number of merging ramps	0.84	1.00	0	4
Diverge	Number of diverging ramps	1.75	2.27	0	17
Inter	Number of intersections	1.90	2.37	0	16
Gradient	Average segment gradient (10^{-2})	0.04	2.74	-11	11
Curvature	Average segment curvature	21.9	17.5	0	85
LCO	Lane changing opportunity	2.43	1.61	0	7.85
Median	Presence of median barrier: yes = 1, no = 0	0.70	0.46	0	1
BS	Presence of bus stop: yes = 1, no = 0	0.64	0.48	0	1
Shoulder	Presence of hard shoulder: yes = 1, no = 0	0.13	0.34	0	1
Park	Presence of on-street parking: yes = 1, no = 0	0.51	0.49	0	1
Rainfall	Annual precipitation (mm)	2279	565	761	3215

1 **Table 2** Model comparison

	Injury severity	Model	Training mean absolute deviation	Testing mean absolute deviation	Number of input nodes	Number of hidden nodes
1	Slight injury	Multivariate Poisson-lognormal	2.87	3.03	—	—
		Trained neural network	2.85	2.99	14	10
		Optimized neural network	2.71	2.93	9	4
	Fatality or serious injury	Multivariate Poisson-lognormal	1.17	1.00	—	—
		Trained neural network	1.09	1.07	14	10
		Optimized neural network	1.08	1.03	8	4
2	Slight injury	Multivariate Poisson-lognormal	2.85	3.00	—	—
		Trained neural network	2.77	2.95	14	10
		Optimized neural network	2.76	2.88	7	4
	Fatality or serious injury	Multivariate Poisson-lognormal	1.09	1.30	—	—
		Trained neural network	1.06	1.26	14	10
		Optimized neural network	1.03	1.25	8	4
3	Slight injury	Multivariate Poisson-lognormal	2.89	2.86	—	—
		Trained neural network	2.84	2.80	14	10
		Optimized neural network	2.63	2.71	11	4
	Fatality or serious injury	Multivariate Poisson-lognormal	1.13	1.15	—	—
		Trained neural network	1.08	1.14	14	10
		Optimized neural network	1.06	1.13	10	4

2

Table 2 (continued) Model comparison

	Slight injury	Multivariate Poisson-lognormal	2.87	2.95	—	—
		Trained neural network	2.78	2.85	14	10
		Optimized neural network	2.71	2.76	9	3
	Fatality or serious injury	Multivariate Poisson-lognormal	1.14	1.13	—	—
		Trained neural network	1.09	1.10	14	10
		Optimized neural network	1.06	1.07	7	4
	Slight injury	Multivariate Poisson-lognormal	2.90	2.82	—	—
		Trained neural network	2.87	2.76	14	10
		Optimized neural network	2.75	2.70	11	4
	Fatality or serious injury	Multivariate Poisson-lognormal	1.12	1.21	—	—
		Trained neural network	1.07	1.12	14	10
		Optimized neural network	1.02	1.08	11	6

Table 3 Rule conditions for slight injury crashes

Rule	Condition[a]
1	v_1 <-1.498 & -0.87≤ v_2 ≤0.87 & -1.042≤ v_3 ≤1.042 & v_4 <-0.52
2	v_1 >1.498 & -0.87≤ v_2 ≤0.87 & -1.042≤v_3 ≤1.042 & v_4 <-0.52
3	v_1 >1.498 & v_2 >0.87 & -1.042≤v_3 ≤1.042 & v_4 <-0.52
4	v_1 <-1.498 & -0.87≤ v_2 ≤0.87 & v_3 >1.042 & v_4 <-0.52
5	v_1 >1.498 & -0.87≤ v_2 ≤0.87 & v_3 >1.042 & v_4 <-0.52
6	v_1 >1.498 & v_2 >0.87 & v_3 >1.042 & v_4 <-0.52
7	v_1 <-1.498 & v_2 >0.87 & -1.042≤ v_3 ≤1.042 & -0.52≤ v_4 ≤0.52
8	v_1 >1.498 & v_2 >0.87 & -1.042≤ v_3 ≤1.042 & -0.52≤ v_4 ≤0.52
9	v_1 >1.498 & -0.87≤ v_2 ≤0.87 & v_3 >1.042 & -0.52≤ v_4 ≤0.52
10	v_1 >1.498 & v_2 >0.87 & v_3 >1.042 & -0.52≤ v_4 ≤0.52
11	v_1 >1.498 & v_2 >0.87 & -1.042≤ v_3 ≤1.042 & v_4 >0.52

[a] $v_1 = -0.819 + 0.344AADT - 1.061Length + 0.159SL - 0.357Median + 0.032BS - 0.443Gradient + 0.076LCO + 0.760Rainfall$,

$v_2 = 0.668 + 1.555AADT - 0.532Length - 0.019SL - 0.179Median + 0.163BS - 0.134Gradient + 0.238LCO - 0.029Rainfall$,

$v_3 = 1.809 - 1.502AADT - 0.763Length - 0.030SL + 0.736Median - 0.305BS + 0.635Gradient - 0.878LCO - 0.323Rainfall$,

$v_4 = -0.504 + 1.271AADT - 0.647Length + 0.285SL - 0.364Median + 0.174BS - 0.177Gradient + 0.504LCO - 0.440Rainfall$.

Table 4 Rule conditions for fatality or serious injury crashes

Rule	Condition[a]
12	v_1 <-1.148 & -0.407≤ v_2 ≤0.407 & v_3 <-0.746 & v_4 <-0.139
13	v_1 >1.148 & -0.407≤ v_2 ≤0.407 & v_3 <-0.746 & v_4 <-0.139
14	v_1 <-1.148 & -0.407≤ v_2 ≤0.407 & -0.746≤ v_3 ≤0.746 & v_4 <-0.139
15	v_1 >1.148 & -0.407≤ v_2 ≤0.407 & -0.746≤ v_3 ≤0.746 & v_4 <-0.139
16	v_1 <-1.148 & v_2 >0.407 & -0.746≤ v_3 ≤0.746 & v_4 <-0.139
17	v_1 >1.148 & v_2 >0.407 & -0.746≤ v_3 ≤0.746 & v_4 <-0.139
18	v_1 >1.148 & -0.407≤ v_2 ≤0.407 & v_3 >0.746 & v_4 <-0.139
19	v_1 <-1.148 & v_2 >0.407 & v_3 >0.746 & v_4 <-0.139
20	v_1 >1.148 & v_2 >0.407 & v_3 >0.746 & v_4 <-0.139
21	v_1 >1.148 & -0.407≤ v_2 ≤0.407 & v_3 <-0.746 & -0.139≤ v_4 ≤0.139
22	v_1 >1.148 & -0.407≤ v_2 ≤0.407 & v_3 <-0.746 & v_4 >0.139

[a] $v_1 = -1.126 + 0.789AADT + 1.082Length + 0.923Diverge - 0.353Median - 0.066BS - 0.281Gradient - 0.317Rainfall$,

$v_2 = 0.348 + 0.352AADT + 0.127Length - 0.6Diverge - 0.483Median + 0.473BS - 0.06Gradient - 0.111Rainfall$,

$v_3 = 0.177 + 0.966AADT - 1.36Length - 0.383Diverge - 1.079Median + 0.751BS - 0.034Gradient + 0.188Rainfall$,

$v_4 = -0.69 - 0.424AADT + 1.047Length - 0.496Diverge + 0.084Median + 0.06BS - 0.126Gradient - 0.587Rainfall$.

Table 5 Rule consequences for slight injury crashes

Rule	Coefficient of the variable in the consequence (linear function)								
	Constant	AADT	Length	SL	Median	BS	Gradient	LCO	Rainfall
1	-0.618	2.110	0.746	-0.105	-0.626	0.347	-0.529	0.830	0.257
2	-0.830	2.110	0.746	-0.105	-0.626	0.347	-0.529	0.830	0.257
3	-0.751	1.322	1.015	-0.095	-0.535	0.265	-0.461	0.710	0.242
4	-0.038	0.975	0.170	-0.127	-0.069	0.117	-0.049	0.167	0.012
5	-0.250	0.975	0.170	-0.127	-0.069	0.117	-0.049	0.167	0.012
6	-0.172	0.188	0.439	-0.117	0.021	0.034	0.019	0.046	0.002
7	-0.543	1.055	1.152	-0.155	-0.458	0.228	-0.424	0.604	0.335
8	-0.755	1.055	1.152	-0.155	-0.458	0.228	-0.424	0.604	0.335
9	-0.254	0.707	0.306	-0.187	0.007	0.080	-0.012	0.061	0.105
10	-0.175	0.080	0.576	-0.177	-0.098	0.003	0.056	-0.060	0.091
11	-0.971	1.322	1.015	-0.095	-0.535	0.265	-0.461	0.710	0.242

Table 6 Rule consequences for fatality or serious injury crashes

Rule	Coefficient of the variable in the consequence (linear function)							
	Constant	AADT	Length	Diverge	Median	BS	Gradient	Rainfall
12	0.336	0.546	0.174	0.049	-0.366	0.225	-0.071	0.038
13	0.994	0.546	0.174	0.049	-0.366	0.225	-0.071	0.038
14	0.0667	0.264	0.570	0.161	-0.051	0.006	-0.061	-0.017
15	0.725	0.264	0.570	0.161	-0.051	0.006	-0.061	-0.017
16	0.080	0.186	0.542	0.294	0.056	-0.099	-0.048	0.008
17	0.738	0.186	0.542	0.294	0.056	-0.099	-0.048	0.008
18	0.559	0.546	0.174	0.049	-0.366	0.225	-0.071	0.038
19	-0.086	0.467	0.146	0.182	-0.258	0.120	-0.058	0.063
20	0.572	0.467	0.146	0.182	-0.258	0.120	-0.058	0.063
21	1.913	1.253	-1.574	0.876	-0.505	0.125	0.138	1.017
22	0.531	0.546	0.174	0.049	-0.366	0.225	-0.071	0.038

Table 7 Parameter estimation in the multivariate Poisson-lognormal model

Variable	Slight injury			Fatality or serious injury		
	Mean	S.D.	95 % Credible interval	Mean	S.D.	95 % Credible interval
Constant	**1.136**	**0.201**	**(0.774, 1.481)**	-0.474	0.307	(-1.052, 0.126)
ln(AADT)	**0.563**	**0.035**	**(0.496, 0.633)**	**0.334**	**0.056**	**(0.228, 0.446)**
ln(Length)	**0.557**	**0.030**	**(0.498, 0.616)**	**0.670**	**0.049**	**(0.574, 0.765)**
SL	**-0.027**	**0.002**	**(-0.034, -0.021)**	**-0.017**	**0.004**	**(-0.026, -0.009)**
Merge	-0.046	0.024	(-0.093, 0.001)	-0.064	0.036	(-0.135, 0.006)
Diverge	**0.023**	**0.012**	**(0.0003, 0.046)**	**0.057**	**0.017**	**(0.023, 0.090)**
Inter	0.015	0.012	(-0.010, 0.040)	-0.010	0.019	(-0.047, 0.026)
Median	**-0.185**	**0.068**	**(-0.313, -0.050)**	**-0.316**	**0.111**	**(-0.533, -0.099)**
BS	**0.384**	**0.054**	**(0.282, 0.490)**	**0.300**	**0.085**	**(0.135, 0.464)**
Gradient	**-1.559**	**0.771**	**(-3.069, -0.058)**	**-2.53**	**1.16**	**(-4.816, -0.253)**
Curvature	-0.002	0.001	(-0.007, 0.004)	-0.001	0.002	(-0.007, 0.006)
LCO	**0.104**	**0.017**	**(0.071, 0.136)**	**0.149**	**0.026**	**(0.099, 0.201)**
Width	0.006	0.035	(-0.064, 0.075)	0.089	0.056	(-0.020, 0.197)
Rainfall	**0.083**	**0.037**	**(0.015, 0.156)**	0.062	0.056	(-0.045, 0.172)

The values in bold are those significantly positive or negative with the 95 % credible intervals bounded away from zero.

Table 8 Hyper-parameter estimation in multivariate Poisson-lognormal

Hyper-parameter	Mean	S.D.	95 % Credible interval	
			2.5 %	97.5 %
σ_{11}	**0.228**	**0.020**	**0.191**	**0.269**
$\sigma_{21}(=\sigma_{12})$	**0.199**	**0.022**	**0.198**	**0.243**
σ_{22}	**0.299**	**0.042**	**0.222**	**0.388**
ρ [a]	**0.763**	**0.047**	**0.667**	**0.847**

The values in bold are those significantly positive or negative with the 95 % credible intervals bounded away from zero.

a: $\rho = \sigma_{12} \Big/ \sqrt{\sigma_{11}\sigma_{22}}$.