Scientia Iranica A (2014) 21(2), 263–275

Sharif University of Technology
Scientia Iranica
Transactions A: Civil Engineering
www.scientiairanica.com

Prediction of roadway accident frequencies: Count regressions versus machine learning models

H. Nassiri[a], P. Najaf[b,*] and A. Mohamadian Amiri[a]

a. *Department of Civil Engineering, Sharif University of Technology, Tehran, P.O. Box 11155-8639, Iran.*
b. *Department of Civil and Environmental Engineering, University of North Carolina at Charlotte, NC, USA.*

Received 10 August 2012; received in revised form 9 April 2013; accepted 22 July 2013

KEYWORDS
Accident frequency prediction;
Negative binomial regression;
Zero inflated negative binomial regression;
Support vector machine;
Back-propagation neural network.

Abstract. Prediction of accident frequency based on traffic and roadway characteristics has been a very significant tool in the field of traffic management. The accident frequencies on 185 roadway segments of the city of Mashhad, Iran, for the year 2007, were used to develop accident prediction models. Negative Binomial Regression, Zero Inflated Negative Binomial Regression, Support Vector Machine and Back-Propagation Neural Network models were used to fit the accident data. Both fitting and predicting abilities of the models were evaluated through computing error values.

Results show that the NBR model is the most effective model for predicting the number of accidents because of its low prediction and fitting error values. Although the BPNN model has high fitting capability, it does not have the prediction ability of the NBR model. Furthermore, the NBR is easily able to develop and interpret the role of effective variables, in comparison with machine learning models which have a black-box form. Marginal effect values for the NBR and ZINBR models, and sensitivity analysis of the SVM and BPNN models, reveal that Volume to Capacity ratio (V/C), Vehicle-Kilometers Travelled (VKT) and roadway width are the most significant variables. An increase in V/C and roadway width will decrease the number of accidents, however, an increase in VKT and permission to park on the right lane of the roadway can increase the crash frequency.

1. Introduction

Rapid population growth and urbanization result in a significant increase in the number of vehicles and, thus, traffic accidents. The issue of accident prediction methods has recently received considerable attention in the field of traffic management. The importance of accident prediction lies in its effect on reducing accidents, injuries, financial losses and delays, as direct costs. In addition, energy waste, missing work days, and economic and psychological consequences, are some of the indirect costs associated with traffic accidents. Iran's reported number of traffic deaths in 2009 was 22,918, and the death rate per 100,000 persons was 35.8, consequently ranking Iran as the 11th most unsafe country in the world [1]. Therefore, evaluation of accident prediction modeling approaches can be beneficial in advancing traffic safety.

Traffic accidents take place due to a number of factors, such as the road, the environment, the driver and the vehicle, and their interrelationship. Therefore, it is important to develop traffic accident prediction models using these parameters to study their effect on traffic safety. The most common

*. *Corresponding author. Tel.: +1 704 641 8468*
E-mail addresses: nassiri@sharif.edu (H. Nassiri); pnajaf@uncc.edu (P. Najaf); amir_mohamadian@alum.sharif.edu (A. Mohamadian Amiri)

crash indicators used are the number of crashes per year (i.e. crash frequency) [2,3], and the number of crashes per million vehicle-kilometers or per mile (i.e. crash rate) [4,5]. Most researchers have focused on determining the relationship between crashes and highway traffic volume, either at the aggregated levels using Annual Average Daily Traffic (AADT) or disaggregated levels by Hourly Volumes (HV) [6,7]. Other studies have further examined the safety of freeway segments as a function of traffic congestion using Volume to Capacity ratio (V/C) or the Congestion Index as independent variables [8]. Crash frequency and aggregate traffic flow characteristics are considered dependent and independent variables in this research, respectively.

The objective of this study is to evaluate the relationship between the number of accidents and independent variables such as V/C, average speed of traffic flow, Vehicle-Kilometers Travelled (VKT), whether parking is permitted on the right lane of the roadway, presence or lack of median, and the width of the roadway. This evaluation is implemented by developing two types of model; Count Regression and Machine Learning, to predict the number of crashes. Negative Binomial Regression (NBR) and Zero Inflated Negative Binomial Regression (ZINBR) models are the most common types of count regression to model accident frequency. Moreover, the Support Vector Machine (SVM) and the Back-Propagation Neural Network (BPNN) are the most well-known machine learning paradigms. In this research, the dataset is divided into two parts (i.e. training set and testing set). The, models are generated based on the training set of samples and tested for the remaining observations. Both fitting and predicting abilities of the models are assessed through comparison of their error values.

2. Literature review

2.1. Count regression models

Count regression models (i.e. generalized linear models) such as Poisson Regression (PR) and NBR are the most common methods for predicting accident frequency. Persaud and Dzbik [9] evaluated the relationships between crash frequency with Average Daily Traffic (ADT) and HV using generalized linear models. They proved that HV data is a more appropriate measure to predict crashes compared to ADT, since it represents the real-time traffic flow condition. Khattak et al. [10] also applied NBR, as well as PR and zero-inflated Poisson regression models, for spatial analysis and the modeling of traffic incidents for strategic planning. Sawalha and Sayed [11] studied the statistical issues of NBR and PR models in traffic safety modeling. They mentioned that avoiding the over-fitting and analysis of outliers are two most important issues when dealing with count regressions. They also proposed a procedure to develop NBR for accident databases as a flowchart.

The NBR model is commonly used in safety analysis, due to the fact that crash data frequently exhibit over-dispersion [12]. Naderan and Shahi [3,13] used NBR to introduce the concept of crash generation using trip generation data and modeled the accident frequency, which can fit the over-dispersed data better than PR. Park and Lord [14] proposed a finite mixture regression model which can fit the nature of over-dispersion in the crash data perfectly and is useful for capturing heterogeneity in crash prediction models. Different types of NBR are applied to different research to fit datasets with different characteristics. Usman et al. [15] calibrated NBR, generalized NBR and ZINBR to model accident frequency considering road surface conditions during snow storms. They concluded that the generalized NBR is more appropriate for capturing heterogeneity in data. However, Sharma and Landge [16] have mentioned that ZINBR is the most appropriate method for modeling the crash frequency of heavy vehicles.

2.2. Machine learning models

In recent years, the Artificial Neural Network (ANN) has been proven to be an efficient and effective method for modeling traffic accidents. There is much research using ANN methods for modeling the classification of injury severities as a discrete dependent variable [17,18], and also to estimate the number of accidents as a countable dependent variable [19]. The ANN method has two main advantages: It is a reliable method for future crash prediction, and there is no need to assume a pre-defined underlying distribution for response variables [20,21]. The main disadvantages of the ANN method are that these models have a black-box form without an analytical basis, and it requires time consuming computational effort to minimize the over-fitting issue, especially when the sample size is small [20]. Chang [21] compared NBR and ANN methods and concluded that ANN models are more efficient in predicting accident frequency. Since there are different types of ANN model with different ranges of characteristics (e.g. different functions, number of layers and number of neurons), there is a need to find out the best type of ANNs for future studies [21].

Recently, application of the SVM model in transportation engineering has been growing. Li et al. [22] developed SVM and NBR models to predict accident frequency. They proved that the SVM model has a higher prediction ability compared to NBR, and it does not over-fit the data. They also mentioned that SVM has an almost similar performance to the BPNN model.

Table 1. Summery descriptions and statistics of variables.

Variable	Description	Unit	Min	Mean	Max	Std Dev.
Y	Crash frequency	#	0	4	45	8.11
V/C	Volume to capacity ratio	ratio	0.03	0.52	2.09	0.42
Speed	Average speed of traffic flow	km/hr	5.84	38.84	68.17	15.24
Parking	1 if parking is permitted on the right lane of the roadway, 0 otherwise	binary	0	0.48	1	0.50
Width	Width of roadway	meter	3.25	14.95	27.6	5.55
Median	1 if roadway has median, 0 otherwise	binary	0	0.35	1	0.48
ln(VKT)	ln(vehicle-kilometer travelled)	veh.km/hr	6.80	13.24	16.96	1.35

2.3. Related parameters

There are some studies about the impact of congestion on accident frequency. Zhou and Sisiopiku [23] studied the correlation between hourly accident rates and average HV per capacity (i.e. V/C). They found a U-shaped relationship between these two variables. Lord et al. [24] considered traffic volume, density and V/C for rural and urban freeway segments to study the relationship between crashes and hourly traffic flow characteristics. They concluded that the volume should not be the only important factor in developing crash prediction models, as there are other important variables, such as density and V/C. Shefer and Rietveld [25] proved that by an increase in density, the number of cars increases. "However, when density becomes so high that speeds are influenced negatively, the number of accidents will decrease [25]". Some researchers use the output of some planning software packages to develop accident prediction model, as well as this research. Hadayeghi et al. [26] used EMME/2 software package to extract traffic flow data (for traffic assignment of the morning peak-period) to develop macro-level accident prediction models.

3. Data description

The studied dataset for this research is based on traffic and accident data from 185 roadway segments of the city of Mashhad, Iran, in the year 2007. This dataset is obtained from the Mashhad Traffic and Transportation Organization of Mashhad Municipality [27]. Three types of variables are used in this dataset including accident frequencies, traffic flow and roadway characteristics. EMME/2 software is used to extract morning peak hour traffic information for the above dataset. In order to validate the EMME/2 estimates, they were compared with actual traffic data at some random segments.

A total of 740 accidents on 185 roadway segments of Mashhad during the morning peak hour (i.e. 7:00-8:00 A.M.) in the year 2007, are considered as the dependent variable. Considered traffic flow independent variables are V/C, average speed of vehicles and VKT. Whether parking is permitted on the right lane of the roadway, the presence or lack of median as a divider, and the width of the roadway are other independent variables regarding roadway characteristics. Table 1 describes the studied parameters.

The results from Table 1 show that the mean and variance of sample accident frequencies are 4 and 65.78, respectively. This indicates that data is over-dispersed, because the variance is significantly larger than the mean. Since one of the most important requirements of Poisson distribution is that the mean of the count data should be equal to its variance, PR cannot fit the dataset properly [28]. These cases are usually modeled by NBR [13].

4. Methodology

NBR, ZINBR, SVM and BPNN are described in this part briefly.

4.1. Negative binomial regression model

If the variance of the countable dependent variable is significantly greater than its mean, the data is considered as over-dispersed. The NBR model is generally applied to fit this type of data [13,28]. The difference between variance and mean is presented by [13]:

$$V[y] = \lambda + \alpha\lambda^2, \tag{1}$$

where $V[y]$ and λ are the estimated variance and mean of crash frequency, and α is the over-dispersion parameter [28].

Eq. (2) shows the general form of the NBR model to predict accident frequency:

$$\lambda_i = \exp(\beta X_i + \varepsilon_i), \tag{2}$$

where exp (ε_i) is a gamma-distributed error with a mean and variance of 1 and α^2 respectively. The general form of Negative Binomial distribution is [28]:

$$P(y_i) = \frac{\Gamma[(1/\alpha) + y_i]}{\Gamma(1/\alpha)\lambda_i}\left(\frac{1/\alpha}{(1/\alpha) + \lambda_i}\right)^{1/\alpha} \left(\frac{\lambda_i}{(1/\alpha) + \lambda_i}\right)^{y_i}, \tag{3}$$

where $\Gamma(.)$ is a gamma function. The coefficients can be estimated using the maximum likelihood approach.

4.2. Zero inflated negative binomial regression model

The reason for using zero inflated count models for accident frequency data is that they can fit excess zeros of the accident data [29]. For sample i, there are two process with a probability of φ_i and 1 -φ_i, respectively. The first process generates only zero counts, and the second one generates count values using a NBR model.

$$y_i \sim \begin{cases} 0 & \text{with probability} \quad \varphi_i \\ g(y_i|X_i) & \text{with probability} \quad 1-\varphi_i \end{cases} \tag{4}$$

The probability of $\{Y_i = y_i|X_i\}$ is:

$$P(Y_i = y_i|X_i, Z_i) =$$

$$\begin{cases} \varphi(\gamma' Z_i) + (1-\varphi(\gamma' Z_i))\, g(0|X_i) & \text{if} \quad y_i = 0 \\ (1-\varphi(\gamma' Z_i))\, g(y_i|X_i) & \text{if} \quad y_i > 0 \end{cases} \tag{5}$$

The probability φ_i depends on the characteristics of sample i. So φ_i is written as a function of $Z_i'\gamma$, where Z_i' is the vector of zero inflated covariates and γ is the vector of zero inflated coefficients to be estimated. The function, F, that produces $Z_i'\gamma$ as a scalar value is called the zero inflated link function. The logistic function is considered a zero inflated link function in this research.

4.3. Support vector machine

This section is briefly extracted from Basak et al. [30], and there is more detail in their report for interested readers. SVM, as a supervised learning method, is useful for recognizing patterns in data. It can produce a set of hyperplanes in a high or infinite dimensional space to develop classification or regression models [30]. In this study, ε-SVM is utilized for regression purpose. The training data is considered as $\{(x_1, y_1), ..., (x_l, y_l)\} \subset \mathfrak{X} \times \mathfrak{R}$, when $\mathfrak{X}$ denotes the space of input patterns, for instance, $\mathfrak{R}^d$. The final goal of ε-SVM regression, is to find a function, $f(x)$, that has most ε deviation from the actual training data, y_i, which should be as flat as possible. The linear function, f, would be described as follows [30]:

$$f(x) = \langle \omega, x\rangle + b \quad \text{with} \quad \omega \in \mathfrak{X}, b \in \mathfrak{R}, \tag{6}$$

where $\langle ., . \rangle$ denotes the dot product in $\mathfrak{X}$ Here, flatness means small ω. Therefore, it is required to minimize the Euclidean norm, as follows [30]:

$$\text{Minimize} \quad \frac{1}{2}||\omega||^2 + C\Sigma_{i=1}^{l}(\xi_i + \xi_i^*)$$

$$\text{Subject to} \begin{cases} y_i - \langle \omega, x_i\rangle - b \le \varepsilon + \xi_i \\ \langle \omega, x_i\rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{7}$$

where ξ_i and ξ_i^* are slack variables, and constant $C > 0$ is a regularization parameter. To extend the SVM to nonlinear functions, the dual formulation is used, as follows [30]:

$$\text{Maximize}\{-\frac{1}{2}\Sigma_{i,j=1}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle$$

$$- \varepsilon\Sigma_{i=1}^{l}(\alpha_i + \alpha_i^*) + \Sigma_{i=1}^{l} y_i(\alpha_i - \alpha_i^*)\}$$

$$\text{Subject to } \Sigma_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C] \tag{8}$$

where α_i and α_i^* are Lagrange multipliers. Finally, the regression function has the following form [30]:

$$f(X) = \Sigma_{i=1}^{l}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b, \tag{9}$$

and b can be computed as follows [30]:

$$b = y_i - \langle \omega, x_i\rangle - \varepsilon \quad \text{for} \quad \alpha_i \in (0, C),$$

$$b = y_i - \langle \omega, x_i\rangle - \varepsilon \quad \text{for} \quad \alpha_i^* \in (0, C). \tag{10}$$

The SVM algorithm can be changed to nonlinear form by preprocessing the training values, x_i, by a map, $\phi : X \to \mathfrak{J}$, into some feature space, $\mathfrak{J}$ and finally applying the standard SVM algorithm. The expansion of Eq. (9) becomes [30]:

$$\omega = \Sigma_{i=1}^{l}(\alpha_i - \alpha_i^*)\phi(x_i).$$

Therefore [30]:

$$f(x) = \Sigma_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_i, x) + b. \tag{11}$$

ω is no longer explicitly given in the non-linear form. The optimization problem in non-linear form tries to find the flattest function in feature space, not in input space [30], where $k(x_i, x)$ is defined as kernel function. There are several types of kernel function, including Linear, Polynomial, Radial Basis and Sigmoid Kernel functions. Generally, prediction performance is performed by the Gaussian Radial Basis Function (GRBF), which is defined as follows [31]:

$$K(x_i, x) = \exp(-\gamma||x_i - x||^2), \tag{12}$$

where γ is a kernel parameter. "Optimization problem is a convex quadratic programming problem, which means that once the kernel function and the input parameters, C, ε are determined, there will be a unique solution for ω and b [31]." In this research, an ε-SVM model with the GRBF kernel is utilized, thus, parameters C, ε and γ need to be determined.

4.4. Back-propagation neural network

This part is briefly extracted from HayKin [32] and Kim et al. [33], and there is more detail in their reports for interested readers. The BPNN structure contains an input layer, a hidden layer, an output layer and connections between them. The BPNN modeling process includes forward and backward phases in its learning algorithm, based on an iterative generalized delta rule with a gradient descent of error. The final objective of this process is to minimize the total error by modification of connection weights.

W_{ji} and W_{kj} are the initial values for connection weights, and θ_j and θ_k are initial biases which must be assumed. At first, in the input layer, the input values, net_{pi}, are activated on the neurons. Then, training and testing values are prepared. Calculation of the "input values of a hidden layer, j, net_{pj}, using the output values of an input layer, i, O_{pi}, connection weight, W_{ji}, and baises, θ_j, between an input layer, i, and a hidden layer, j, is the next step. Finally, the output values of a hidden layer, j, O_{pj}, are derived from net_{pj} [33]:

$$\text{net}_{pj} = \Sigma_i W_{ji} O_{pi} + \theta_j, \tag{13}$$

$$O_{pj} = f_j(\text{net}_{pj}), \tag{14}$$

where $f(.)$ is an activation function. In this study, the hyperbolic tangent function is used as the activation function, which has a greater range than other common functions.

"Input values of an output layer, k, net_{pk}, are computed using the output values of a hidden layer, j, O_{pi}, connection weight, W_{kj}, and biases, θ_k, between a hidden layer, j, and an output layer, k. Then, the output values of an output layer, k, O_{pk}, are derived from net_{pk} [33]":

$$\text{net}_{pk} = \Sigma_j W_{kj} O_{pj} + \theta_k, \tag{15}$$

$$O_{pk} = f_k(\text{net}_{pk}). \tag{16}$$

To modify the connection weights and biases based on the generalized delta rule, the error at the output neurons is propagated backward to the hidden layer, and then to the input neurons. These steps are from the hidden layer to the output layer's neurons [33]:

$$\Delta W_{kj} = \eta \delta_k O_{pj} \quad \text{and} \quad \Delta B_k = \eta \delta_k, \tag{17}$$

where $\delta_k = (T_k - O_{pk}) f'(\text{net}_{pj})$ and η = the learning rate; and from input layer to hidden layer's neurons:

$$\Delta W_{ji} = \eta \delta_j \text{net}_{pi} \text{ and } \Delta B_j = \eta \partial_j, \tag{18}$$

where:

$$\delta_j = W_{kj} \delta_k f'(\text{net}_{pj}).$$

The error, E between the calculated value, O_{pk} and the desired value T_k is defined as [33]:

$$E = \frac{1}{2} \Sigma_{k=1} (O_{pk} - T_k)^2. \tag{19}$$

This procedure should be repeated until error E goes below a target value.

5. Modeling results

The dataset is randomly divided into two categories. Seventy percent of the roadway segments' observations (i.e. 130 roadway segments) are utilized to calibrate the models and evaluate the models' fitting ability (i.e. first part or training set of samples). The remaining thirty percent of roadway segments' observations (i.e. 55 roadway segments) are employed to compare the prediction capability of the models and compute error values (i.e. second part or testing set of samples). This section is comprised of three parts: Development of NBR, ZINBR, SVM and BPNN models using the training set of samples (Model Development), evaluation of the models' fitting ability based on the training set of samples (Fitting), and evaluation of the models' prediction ability based on the testing set of samples (Prediction).

5.1. Model development

In this section, the models are generated using the first part of the observations.

5.1.1. NBR

The NBR model is developed by NLOGIT 4 software [34]. The coefficients and statistical characteristics of the model are presented in Table 2. V/C, parking, width and ln(VKT) are significant variables. Speed and median are insignificant variables which are omitted from the final NBR model. The model reveals a proper quality in statistical tests (Significance Level=0.00 and Pseudo R-squared=0.37). Also, the negligible P-value of the dispersion parameter (alpha) in the NBR model proves that the model has a reasonable statistical performance in comparison with other types of count regression model, such as PR.

5.1.2. ZINBR

Since accident frequency for 50 percent of the first part of our observations (i.e. training set of samples) is zero, ZINBR may be an appropriate model to use. The ZINBR model is developed by NLOGIT 4 software [34] for the first part of the observations, presented in Table 3. According to the methodology section, the ZINBR model has two separate parts. The right side of the table represents a binary logit (i.e. logistic function) to model the probability of accident occurrence on roadway segments, and the left side of the table represents a simple NBR. A logistic function

Table 2. Negative binomial regression model.

Variables	Coefficient	t-statistic	P-value	Marginal effect
Constant	-10.3262	-16.91	0.0000	----
V/C	-1.29288	-4.82	0.0000	-4.96
Parking	0.69727	3.93	0.0001	2.67
Width	-0.06091	-3.22	0.0013	-0.23
ln(VKT)	0.93240	18.69	0.0000	3.58
Alpha*	0.83918	5.39	0.0000	----
Goodness of fit				
Number of observations				130
Log likelihood				-274.69
Restricted log likelihood				-440.88
Chi squared				332.38
Significance level				0.0000
Pseudo R-squared				0.3769

* Dispersion parameter for the NBR model

Table 3. Zero inflated negative binomial regression model.

Zero inflation model					Negative binomial regression				
Variables	Coeff.	t-statistic	P-value	Marginal effect	Variables	Coeff.	t-statistic	P-value	Marginal effect
Constant	0.14452	0.31	0.758	-	Constant	-5.95834	-8.591	0.000	-
V/C	-2.30448	-1.981	0.047	-1.43	V/C	-1.04252	-3.316	0.001	-1.43
-	-	-	-	-	ln(VKT)	0.59052	11.220	0.000	1.92
-	-	-	-	-	Alpha*	1.26817	4.342	0.000	-
Goodness of fit									
Number of observations									130
Number of zero observations									65
Log likelihood									-262.96
Restricted log likelihood									-478.07
Significance level									0.0000
Pseudo R-squared									0.4491
Vuong test of ZINBR versus NBR									1.22 (Pr>z=0.1107)

*: Dispersion parameter for the NBR model.

is applied as the splitting distribution function for the first part of the ZINBR model. In the generated ZINBR model, V/C and ln(VKT) are the most significant variables. However, the calibrated ZINBR model is statistically significant (Significance Level = 0.00 and Pseudo R-squared = 0.44), but it does not show better statistical performance than the generated NBR model. The value of the Vuong test is 1.22 (less than 1.96), which demonstrates the lower fitting ability of the ZINBR in comparison with the NBR model.

Since the ZINBR model contains two separate parts, the selection of significant variables is very critical. Indeed, if one part of the ZINBR model cannot be calibrated well (e.g. due to the lack of significant variables), error values for the final prediction of the ZINBR model will increase. In this research, the developed ZINBR model does not show high fitting and prediction ability for this reason.

5.1.3. SVM

The winSVM software is applied to generate the SVM models [35]. As mentioned before, C, ε and γ parameters should be determined to develop the ε-SVM model with the radial basis kernel function. Several models based on different values of C, ε and γ are adapted to determine the best model, which can strongly fit the first part of the observations (i.e. training set of samples). A number of these models are presented in

Table 4. Some ε-SVM models based on different values of C, ε and γ.

C	ε	γ	MSE	C	ε	γ	MSE
100000	0.00001	5	30.83	10000	1	2	31.87
0.01	1	10	30.89	1000	0.1	1	35.35
0.1	0.01	10	30.89	10	0.1	0.5	38.04
0.01	0.001	10	30.89	0.01	0.1	0.2	42.94
100	0.01	10	30.89	1000	0.1	0.05	55.56
1000000	1	100	30.98	1000000	0.01	0.01	61.74
100	0.0001	100	30.98	1000	0.00001	0.001	72.24

Table 4 based on different values of C, ε, γ and Mean Squared Error (MSE).

In the SVM model development process, it was observed that the MSE value is strongly dependent on the value of the γ parameter. In other words, the model's sensitivity to variations of γ is greater compared to variations in C and ε values. Consequently, assuming $C = 1$ and $\varepsilon = 1$, the SVM models with different values of γ were developed. The MSE values of these models in fitting the first part of the samples are displayed in Figure 1.

According to this figure, the SVM model with $\gamma = 5$ has the best fitting ability for the training set of observations. To clarify the issue further, this study tries to answer the following question: Does the SVM model with the best fitting ability to the first part of the samples have the best prediction for the second part of the samples as well or not? To address this question, the MSE values are computed in order to predict the second part of the samples. Moreover, these values are illustrated in Figure 1. The lowest prediction error occurs in $\gamma = 0.2$. Therefore, the MSE value for the SVM model with $C = 1$, $\varepsilon = 1$ and $\gamma = 5$ in fitting the training observations is 30.83; furthermore, the MSE is 65.41 in predicting the testing observations (i.e. total MSE is 41.11). On the other hand, the MSE of the SVM model with $C = 1$, $\varepsilon = 1$ and $\gamma = 0.2$ fitting to the training observations is 50.33, and, further, is 42.94 in predicting the second part of the observations (i.e. total MSE is 45.12). Therefore, it can be concluded that the SVM model with the best fitting ability for a part of the samples may not have the best prediction for other samples, and vice versa.

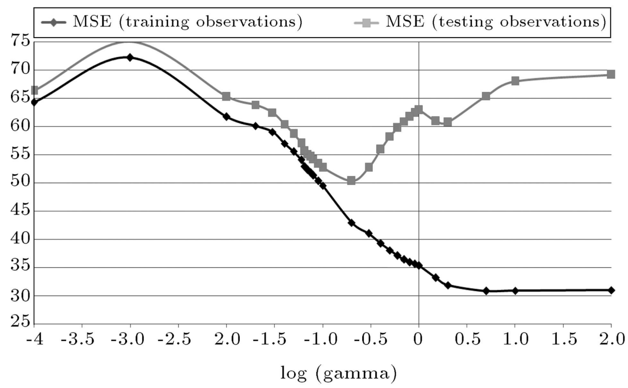

Figure 1. MSEs for different values of γ in the SVM model.

5.1.4. BPNN

The applied software for this part of the research is entitled Neurosolutions 5 [36]. The models in this section are trained for the first part of the samples and tested for the second part. The hyperbolic tangent is assumed to be the activation function for all neurons. The numbers of epochs, the learning rate and the momentum for the training process are set to 1000, 1 and 0.7, respectively. Considering that the number of hidden layers and neurons in each hidden layer can be a wide range of values, a number of different models with one hidden layer are trained in this section. The number of neurons in the hidden layer of these models is varied between 1 and 20. The objective of training different models is to identify which model has the best fitting and prediction abilities. Figure 2 clarifies the MSE of the BPNN models in fitting the first part of the observations and predicting the second part of the samples.

The figure explains that MSE values do not have an explicit correlation with the number of neurons. In addition, it is clear that the model with the best fitting for training observations may not predict testing

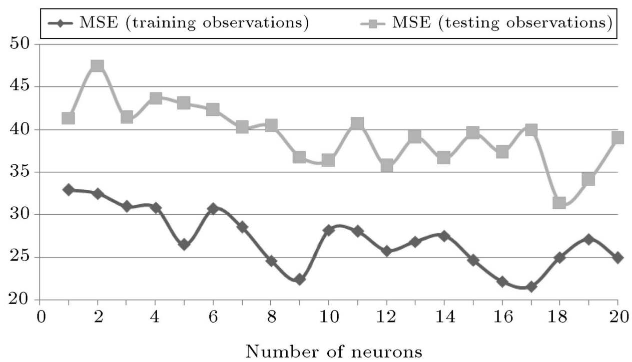

Figure 2. MSEs for different number of neurons in the BPNN model.

observations strongly, the same as the SVM model. The BPNN model with 17 neurons in the hidden layer has the best fitting to the training set of samples (MSE = 21.61. The MSE of this model is 39.98 in predicting the testing set of samples, and its total MSE is 27.07. The BPNN model with 18 neurons in the hidden layer has MSE = 25.00 in fitting and MSE = 31.43 in predicting (i.e. totally 26.91). Therefore, the BPNN model that has the best fitting does not have the best prediction among the trained models.

5.2. Fitting ability

In this section, the fitting ability of the generated models is evaluated comparing their error values. Table 5 shows the following types of errors for the NBR, ZINBR, SVM and BPNN models.

Mean Squared Error (MSE)

$$= \frac{\sum_{(j=1)}^{N}(y_j - \hat{y}_j)^2}{N}$$

Normalized Mean Squared Error (NMSE)

$$= \frac{\text{MSE}}{\text{Var}(y_j)}$$

Mean Absolute Error (MAE)

$$= \frac{\sum_{(j=1)}^{N}|y_j - \hat{y}_j|}{N}$$

Minimum Absolute Error (Min AE)

$$= \min\{|y_j - \hat{y}_j|, j = 1, \cdots, N\}$$

Maximum Absolute Error (Max AE)

$$= \max\{|y_j - \hat{y}_j|, j = 1, \cdots, N\}$$

Root Mean Squared Error (RMSE)

$$= \sqrt{\frac{1}{N}\sum_{(j=1)}^{N}(y_j - \hat{y}_j)^2}$$

These values are computed based on the training set of samples (i.e. first part of observations), which is only for model development and evaluation of the models' fitting strength.

Results prove that the BPNN models have higher fitting capability for the training set of samples. The SVM ($\gamma = 5$) and NBR also have acceptable performance. As mentioned before, the ZINBR does not show an appropriate performance in fitting the first part of the observations, because it is not more statistically significant than the NBR.

5.3. Prediction ability

The ability of the generated models in predicting the second part of the observations (i.e. thirty percent of samples, testing set of samples) is compared based on error values. In addition to the NBR and ZINBR models, the SVM and BPNN models which had the best fitting and the best prediction are presented (NBR, ZINBR, SVM ($\gamma = 5$), SVM ($\gamma = 0.2$), BPNN (17 neurons) and BPNN (18 neurons)). Table 6 shows different error values of the aforementioned calibrated models in predicting the second part of the samples.

The BPNN model with 18 neurons in the hidden layer has the lowest prediction error values in

Table 5. Model's fitting error values.

Errors	NBR	ZINBR	SVM (γ=5)	SVM (γ=0.2)	BPNN (17 neuron)	BPNN (18 neuron)
MSE	35.6019	46.3177	30.8353	42.9443	21.6111	25.0048
NMSE	0.5447	0.7087	0.4717	0.6568	0.3306	0.3826
MAE	3.4354	3.7954	1.4038	2.6136	2.9824	3.1749
Min AE	0.0137	0.0395	0.0000	0.0000	0.0011	0.0075
Max AE	33.7124	37.4552	34.5940	39.5422	20.1335	20.5133
RMSE	5.9667	6.8057	5.5523	6.5519	4.6487	5.0004

Table 6. Model's prediction error values.

Errors	NBR	ZINBR	SVM (γ=5)	SVM (γ=0.2)	BPNN (17 neuron)	BPNN (18 neuron)
MSE	34.5960	54.4839	65.4110	50.3326	39.9888	31.4340
NMSE	0.5087	0.8012	0.9618	0.7401	0.5880	0.4622
MAE	3.8084	3.9275	4.1439	3.8218	4.4427	4.1528
Min AE	0.0845	0.1197	0.0015	0.0151	0.2779	0.1226
Max AE	24.4812	35.2419	39.3179	34.2979	18.7077	14.7672
RMSE	5.8818	7.3813	8.0877	7.0945	6.3236	5.6066

comparison with other generated models (Table 6). However, the problem is that it does not have a very high fitting ability for the first part of the samples (Table 5). In real-life accident modeling studies, there is no information about the model's predictions, where the value of the dependent variable for the second part of the observations will not be available. We just know the values of the dependent variable for the first part of the observations. Therefore, the goal will be to predict the number of accidents for the second part of the observations using the calibrated model, based on the first part of the observations. So, the model with the highest fitting ability should be selected, not the model with the highest prediction ability. In this case, if we want to select the best BPNN model among all generated BPNN models, the BPNN model with 17 neurons which has the highest fitting ability, will definitely be selected. Therefore, the final prediction of the best BPNN model (i.e. BPNN with 17 neurons) will not be as accurate as the NBR model.

Furthermore, Table 6 clarifies that both SVM models do not have high prediction ability in comparison with the NBR and BPNNs. In addition, it is clear that the ZINBR is not more accurate than the NBR in prediction, too.

6. Analysis of parameters

In this section, the significance of the studied variables, as well as their effect on the number of accidents, is analyzed. Furthermore, considering the interpretation of the role of factors on accident frequency, the models are compared.

6.1. NBR and ZINBR

NBR and ZINBR models' variables analysis is described considering the coefficient values, signs of estimated coefficients, significance level of parameters and marginal effect values. The marginal effect values are reported in Tables 2 and 3. For this purpose, partial derivatives of the expected values, with respect to the vector of characteristics, are computed, when effects are averaged over all observations. The impact of each variable on the accident frequency variable is briefly described as follows:

- V/C: This variable is significant in the NBR and both parts of the ZINBR models. By increasing this variable, the number of traffic accidents will decrease, because the V/C has negative coefficients in the models. This means that when traffic congestion increases, the number of accidents will decrease. Since the V/C has a negative coefficient in the first part of the ZINBR model, it is concluded that by increasing this parameter, the probability of observing no accidents on the roadway will increase. The marginal effect of this variable shows that the V/C is one of the most effective factors to predict accident frequency in both NBR and ZINBR models.
- Speed: This variable is not significant in the generated count regression models, NBR and ZINBR. Since speed may have a dual effect, more studies are required to analyze the effect of speed on accident prediction using count regressions. In the next section, the effect of speed on accident frequency will be evaluated considering the generated machine learning models, SVM and BPNN.
- Parking: Parked vehicles on the right lane of the roadway cause a reduction of useful roadway width and an increase in traffic interactions when the vehicles enter or exit from the parking space. Therefore, an increase in the number of accidents will result, according to the coefficient of this variable, in the NBR model. This variable is not significant in the calibrated ZINBR model.
- Width: Since it has a negative coefficient in the generated NBR model, the number of accidents will decrease by increasing the useful width of the roadway. This variable is insignificant in the ZINBR model.
- Median: This variable is not significant in the generated count regression models. In the next section, the effect of the median will be explained based on the generated SVM and BPNN models.
- ln(VKT): This variable possesses significant and positive coefficients in both NBR and ZINBR models. An increase in VKT (higher vehicle volume on the roadway) will increase the frequency of accidents. The marginal effect values show that VKT can be one of the most effective factors in the prediction of traffic accidents.

6.2. SVM and BPNN

Since the BPNN and SVM models work as a black-box, it can be said that they are not able to generate interpretable parameters for each explanatory variable in an interpretable functional form. They have no specific function form. To minimize this problem, in some research, the sensitivity analysis of parameters is suggested [22]. The sensitivity analysis for all parameters consisted of recording the changes in accident frequencies generated from the SVM and BPNN models for different values of V/C, speed, parking, width, median and ln(VKT) variables, within reasonable intervals, while keeping all other variables constant. The sensitivity analysis is performed for one segment, segment number 66, which is randomly selected. Table 7 illustrates the segment characteristics.

Figures 3 and 4 show the sensitivity analysis of the SVM and BPNN models, respectively. As is evident

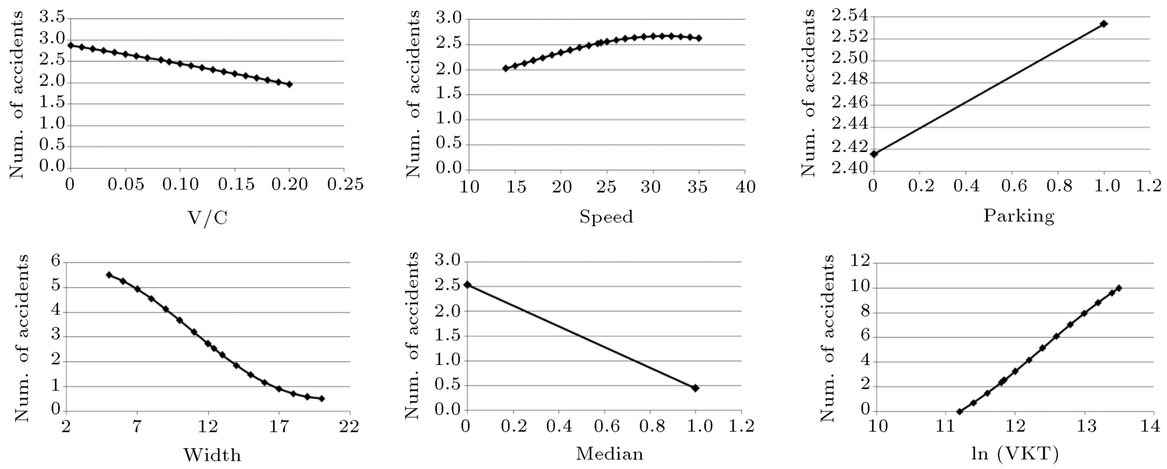

Figure 3. Sensitivity analysis of the SVM model.

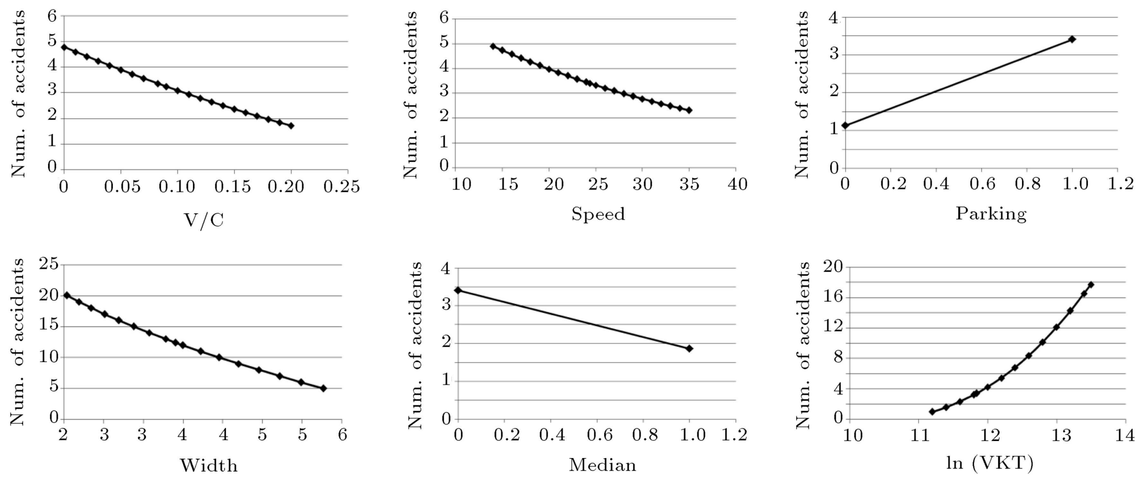

Figure 4. Sensitivity analysis of the BPNN model.

Table 7. Characteristics of analyzed roadway segment.

Observation	# 66	V/C	0.08
Num. of accidents	1	Speed	24.36
Pred. of SVM (γ=5)	1.0036	Parking	1
Pred. of SVM (γ=0.2)	2.5333	Width	12.41
Pred. of BPNN (17 Neuron)	3.2377	Median	0
Pred. of BPNN (18 Neuron)	3.3673	ln(VKT)	11.8240

from the curves, the SVM and BPNN estimations are highly similar for V/C, parking, width, median and ln(VKT) variables. These results are consistent with the NBR and ZINBR results as well. According to the slopes in the figures, V/C, ln(VKT) and width are the most significant variables. Speed and median are not as significant as other variables.

SVM assessment for a speed variable is different from other models. According to the sensitivity analysis of the speed variable in the SVM model, an increase in the average speed of traffic flow will increase accident frequency at lower speeds and, then, will decrease the number of accidents at higher speeds. More studies are required to evaluate the role of speed on accident frequency.

7. Conclusion

The objective of this study is to model the relationship between accident frequency and traffic flow variables, such as V/C, average speed of traffic flow and VKT. Furthermore, whether parking is permitted on the right lane of the roadway, the presence or lack of a roadway median, and the width of the roadway are other studied independent variables. The dataset is based on 185 roadway segments of the city of Mashhad, Iran, in the year 2007, and accidents which occurred during morning peak hour are modeled (i.e. totally 740 crashes). Two types of model are evaluated: count regressions models, including NBR and ZINBR models, and machine learning models, including SVM and BPNN models. The fitting and predicting abilities of these models are evaluated through computing error values.

The results of modeling steps are summarized as follows:

- The ZINBR model has two separate parts: a binary logit (i.e. logistic function) to model the probability of accident occurrence on the links, and a simple NBR to estimate the frequencies. If one of these parts does not show reliable and significant explanatory power, error values will increase in the final prediction. In this research, the calibrated ZINBR model did not show high fitting and prediction ability due to the lack of significant variables in both of its parts. Furthermore, it did not show better statistical performance than the NBR model, because the value of the Vuong test was not in an acceptable range.
- In developing the SVM models, it was observed that the MSE value is strongly dependent on the value of the γ parameter. Among all generated SVM models, the SVM with $\gamma = 5$ had the best fitting for the training set of observations (i.e. first part of the samples), and the lowest error for prediction of the testing set of observations (i.e. second part of the samples) occurred in the SVM model with $\gamma = 0.2$. It can be concluded that the SVM model with the best fitting to a part of the samples, may not have the best prediction of the other part of the samples, and vice versa. SVM models generally showed acceptable fitting ability and low prediction capability. Furthermore, it is time consuming to generate different types of SVM to achieve the best performance. Also, it is difficult to identify the role of independent variables in the SVM because of its black-box form, and difficulties in computing the sensitivity analysis.
- Among different generated BPNNs, the model with 17 neurons showed the best fitting to the training set of samples, and the BPNN model with 18 neurons showed the highest prediction performance. Again, it can be concluded that the BPNN model with the best fitting to a part of the samples may not have the best prediction of the other part of the samples, and vice versa. The BPNN models generally showed the highest fitting ability among all types of studied models. The main problem with the BPNN models was related to their prediction ability. Since, in real-life modeling, we do not have any knowledge about the models' predictions (i.e. because there is no information about the dependent variable), we will have to select the model with the highest fitting ability (BPNN with 17 neurons, in this research). Therefore, the accuracy of the BPNN's final prediction will be unknown (in this research, the prediction accuracy of the BPNN with 17 neurons was less than the prediction accuracy of the NBR model). Furthermore, it is time consuming to generate different types of BPNN to achieve the best performance. Also, it is difficult to identify the role of independent variables in the BPNN because of its black-box form and difficulties in computing the sensitivity analysis.
- The calibrated NBR model showed a proper quality in statistical tests and interpretation of variables. It had more explanatory power than the ZINBR. The generated NBR model, in this research, had a reasonable fitting ability (a little less than BPNNs) and the highest prediction ability (even higher than BPNNs). Furthermore, it is easy to develop, and it is possible to use marginal effect values for variable sensitivity analyses.

According to our findings, the NBR model is suggested to model accident frequency because of its reasonable fitting and high prediction ability. Also, it is easy to develop and interpret the role of different significant variables. Furthermore, it is concluded that V/C and VKT are the most effective factors in predicting the number of accidents; the presence of parked vehicles on the right side of the roadway and roadway width are other significant variables.

Since the models in this research are developed based on a one year dataset of one city in Iran, it is suggested to develop models for different places using longer periods of time to evaluate the temporal and spatial reliability of the models. It is also suggested to model accident frequency in different ranges of average speed to evaluate the role of this variable on the number of accidents.

Abbreviations

NBR	Negative Binomial Regression
ZINBR	Zero Inflated Negative Binomial Regression
SVM	Support Vector Machine
BPNN	Back-Propagation Neural Network
AADT	Annual Average Daily Traffic
HV	Hourly Volume
V/C	Volume to Capacity Ratio
VKT	Vehicle-Kilometers Travelled
ADT	Average Daily Traffic
ANN	Artificial Neural Network
MSE	Mean Squared Error
NMSE	Normalized Mean Squared Error
MAE	Mean Absolute Error
Min AE	Minimum Absolute Error
Max AE	Maximum Absolute Error
RMSE	Root Mean Squared Error

References

1. Road Traffic Deaths Index 2009 Ranking "Global status report on road safety", Accessed July 1, (2011).

http://www.photius.com/rankings/road_traffic_deaths_country_rankings_2009.html.

2. Caliendo, C., Guida, M. and Parisi, A. "A crash-prediction model for multilane roads", *In Accident Analysis and Prevention,* **39**, pp. 657-670 (2007).

3. Naderan, A. and Shahi, J. "Aggregate crash prediction models: Introducing crash generation concept", In *Accident Analysis and Prevention*, **42**(1), pp. 339-336 (2010).

4. Knuiman, M.W., Council, F.M. and Reinfurt, D.W. "Association of median width and highway accident rate (with discussion and closure)", In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1401, Transportation Research Board of the National Academies, Washington, D.C., pp. 70-82 (1993).

5. Zhou, M. and Sisiopiku, V.P. "Relationship between volume to capacity ratio and accident rates", In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1581, Transportation Research Board of the National Academies, Washington, D.C., pp. 47-52 (1995).

6. Mensah, A. and Hauer, E. "Two problems of averaging arising in the estimation of the relationship between accidents and traffic flow", In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1635, Transportation Research Board of the National Academies, Washington, D.C., pp. 37-43 (1998).

7. Martin, J.L. "Relationship between crash rate and hourly traffic flow on interurban motorways", In *Accident Analysis and Prevention*, **34**, pp. 619-629 (2002).

8. Wang, C., Quddus, M.A. and Ison, S.G. "Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England", In *Accident Analysis and Prevention*, **41**, pp. 798-808 (2007).

9. Persaud, B. and Dzbik, L. "Accident prediction models for freeways", In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1401, Transportation Research Board of the National Academies, Washington, D.C., pp. 55-60 (1993).

10. Khattak, A.J., Wang, X. and Zhang, H. "Spatial analysis and modeling of traffic incidents for proactive incident management and strategic planning", In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2178, Transportation Research Board of the National Academies, Washington, D.C., pp. 128-137 (2010).

11. Sawalha, Z. and Sayed, T. "Statistical issues in traffic accident modeling", In *Transportation Research Board, 82th Annual Meeting*, Transportation Research Board of the National Academies, Washington, D.C. (2003).

12. Pei, X., Wong, S.C. and Sze, N.N. "Negative binomial regression model for road accident analysis in Hong Kong", In *Transportation Research Board 89th Annual Meeting*, Transportation Research Board of the National Academies, Washington, D.C. (2010).

13. Naderan, A. and Shahi, J. "Crash generation models: Forecasting crashes in urban areas", In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2148, Transportation Research Board of the National Academies, Washington, D.C., pp. 101-106 (2010).

14. Park, B.J. and Lord, D. "Application of finite mixture models for vehicle crash data analysis", In *Accident Analysis and Prevention*, **4**, pp. 683-691 (2009).

15. Usman, T., Fu, L. and Miranda-Moreno, L.F. "Quantifying safety benefit of winter road maintenance: Accident frequency modeling", In *Accident Analysis and Prevention*, **42**(6), pp. 1878-1887 (2010).

16. Sharma, A.K. and Landge, V.S. "Zero inflated negative binomial for modeling heavy vehicle crash rate on Indian rural highways", In *International Journal of Advances in Engineering & Technology*, **5**(2), pp. 292-301 (2013).

17. Chong, M.M., Abraham, A. and Paprzycki, M. "Traffic accident analysis using decision trees and neural networks", In *IADIS International Conference on Applied Computing*, Portugal, IADIS Press, **2**, pp. 39-42 (2004).

18. Moghaddam, F.R., Afandizadeh, S. and Ziyadi, M. "Prediction of accident severity using artificial neural networks", In *International Journal of Civil Engineering*, **9**(1), pp. 41-48 (2011).

19. Khalili, F., Sheikholeslami, A. and Mahmoudabadi, A. "Variable efficiency appraisal in freeway accidents using artificial neural networks: Case study", In *CICTP*, pp. 2657-2664 (2012).

20. Sando, T., Mussa, R., Sobanjo, J. and Spainhour, L. "Advantages and disadvantages of different crash modeling techniques", In *Journal of Safety Research*, **36**(5), pp. 485-487 (2005).

21. Chang, L.Y. "Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network", In *Safety Science*, **43**(8), pp. 541-557 (2005).

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

23. Zhou, M. and Sisiopiku, V. "On the relationship between volume to capacity ratios in accident rates", In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1581, Transportation Research Board of the National Academies, Washington, D.C., pp. 47-52 (1997).

24. Lord, D., Manar, A. and Vizioli, A. "Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments", In *Accident Analysis and Prevention*, **37**, pp. 185-199 (2006).

25. Shefer, D. and Rietveld, P. "Congestion and safety

on highways: towards an analytical model", In *Urban Studies*, **34**(4), pp. 679-692 (1997).

26. Hadayeghi, A., Shalaby, A.S. and Persaud, B. "Macro-level accident prediction models for evaluating safety of urban transportation systems", In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1840, Transportation Research Board of the National Academies, Washington, D.C., pp. 87-95 (2003).

27. Mashhad Traffic and Transportation Organization, Mashhad Municipality, Mashhad, Iran. http://www.mashadtraffic.ir/en/portal/.

28. Cameron, A.C. and Trivedi, P.K., *Regression Analysis of Count Data*, Cambridge University Press, New York (1998).

29. Erdman, D., Jackson, L. and Sinko, A. "Zero inflated poisson and zero inflated negative binomial models using the COUNTREG procedure", In *SAS Global Forum* (2008).

30. Basak, D., Pal, S. and Patranabis, D.C. "Support vector regression", In *Neural Information Processing - Letters and Reviews*, **11**(10), pp. 203-224 (2007).

31. Zhang, N., Zhang, Y. and Lu, H. "Short-term freeway traffic flow prediction combining seasonal autoregressive integrated moving average and support vector machines", In *Transportation Research Board 90th Annual Meeting*, Transportation Research Board of the National Academies, Washington, D.C. (2011).

32. Haykin, S.O., *Neural Networks: A Comprehensive Foundation*, In Prentice-Hall Inc., New Jersy (1998).

33. Kim, D.K., Lee, J.J., Chang, S.K. and Chang, S.K. "Active vibration control of a structure using neural network techniques", In *Transportation Research Board 86th Annual Meeting*, Transportation Research Board of the National Academies, Washington, D.C. (2007).

34. Greene, W.H. NLOGIT. "Version 4: user's manual", Economic Software, New York (2007). http://www.limdep.com/products/nlogit/.

35. Sewell, M. winSVM., *A Windows Implementation of Support Vector Machine*, (2005). http://www.cs.ucl.ac.uk/staff/M.Sewell/winsvm/.

36. Neurosolutions 5, *The Premier Neural Network Development Environment*, (2010). http://www. neurosolutions.com/index.html.

Biographies

Habibollah Nassiri received his BSc and MSc in Civil Engineering from Auburn University, Auburn, AL, USA in 1979 and 1981, respectively. In 1989, he earned his PhD degree in Civil Engineering specializing in Transportation Planning from Texas A& M University, TX, USA. He joined as a faculty member to Sharif University of Technology, Tehran, Iran, in 1994, where he is currently an Associate Professor. His main research interests include traffic operation and flow theory, traffic safety, traffic monitoring and geometric design. He has published many papers in international journals and conferences.

Pooya Najaf received his BSc in Civil Engineering from Iran University of Science and Technology in 2009. He also earned his MSc in Transportation Engineering and Planning from Sharif University of Technology in 2011. Then he joined University of North Carolina, Charlotte, NC, USA as a research and teaching assistant, where he is currently pursuing his PhD degree. His main research interests include traffic safety, econometrics, traffic simulation and public transportation.

Amir Mohamadian Amiri received his BSc in Civil Engineering from Babol Noshiravani University of Science and Technology, in 2009, and his MSc degree in Highway Engineering from Sharif University of Technology in 2011. He is currently pursuing his PhD degree in Highway Engineering at Iran University of Science and Technology. His main research interests include highway safety, traffic flow theory and traffic operation.