Accident Analysis and Prevention 45 (2012) 478–486

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

# Using support vector machine models for crash injury severity analysis

Zhibin Li[1], Pan Liu*, Wei Wang[2], Chengcheng Xu[3]

*School of Transportation, Southeast University, Si Pai Lou #2, Nanjing 210096, China*

ARTICLE INFO

*Article history:*
Received 20 July 2011
Received in revised form 22 August 2011
Accepted 28 August 2011

*Keywords:*
Support vector machine model
Ordered probit model
Crash severity
Freeway diverge area

ABSTRACT

The study presented in this paper investigated the possibility of using support vector machine (SVM) models for crash injury severity analysis. Based on crash data collected at 326 freeway diverge areas, a SVM model was developed for predicting the injury severity associated with individual crashes. An ordered probit (OP) model was also developed using the same dataset. The research team compared the performance of the SVM model and the OP model. It was found that the SVM model produced better prediction performance for crash injury severity than did the OP model. The percent of correct prediction for the SVM model was found to be 48.8%, which was higher than that produced by the OP model (44.0%). Even though the SVM model may suffer from the multi-class classification problem, it still provides better prediction results for small proportion injury severities than the OP model does.

The research also investigated the potential of using the SVM model for evaluating the impacts of external factors on crash injury severities. The sensitivity analysis results show that the SVM model produced comparable results regarding the impacts of variables on crash injury severity as compared to the OP model. For several variables such as the length of the exit ramp and the shoulder width of the freeway mainline, the results of the SVM model are more reasonable than those of the OP model.



## 1. Introduction

The analysis of crash injury severity is of great interest to many transportation professionals. One of the main objectives of crash injury severity analysis is to understand the relationship between the injury severity of crashes and various contributing factors such as the driver and passenger characteristics, vehicles types, traffic conditions, geometric design characteristics, as well as the collision types of crashes, etc. Such information will help decision makers better understand the impacts of contributing factors on crash injury severity and implement treatments to reduce the severity of crashes.

Injury severity data are generally represented by discrete categories such as fatal, incapacitating injury, non-incapacitating injury, possible injury and property damage only, etc. Traditionally, transportation professionals use statistical models to evaluate the effects of contributing factors on crash injury severity. Among them, the ordered probit (OP) models and their variations are probably the most commonly used modeling techniques (Odonnell and Connor, 1996; Duncan et al., 1998; Kockelman and Kweon, 2002; Abdel-Aty, 2003; Zajac and Ivan, 2003; Abdel-Aty and Abdelwahab, 2004; Lee and Abdel-Aty, 2005; Siddiqui et al., 2006; Yau et al., 2006; Xie et al., 2009; Wang et al., 2011). Some other statistical models have also been proposed for crash injury severity analysis, including the ordered logit model (Odonnell and Connor, 1996), the multinomial logit model (Shankar and Mannering, 1996; Khorashadi et al., 2005; Savolainen and Mannering, 2007), and the logistic regression model (Al-Ghamdi, 2002), etc.

Even though traditional statistical models have been widely used for crash injury severity analysis, they do suffer from some limitations. For example, traditional statistical modeling requires assumptions about the distribution of data and, usually, a linear functional form between dependent and explanatory variables. These assumptions may not always hold true. When basic assumptions of traditional statistical models were violated, erroneous estimations and incorrect inferences could be produced (Mussone et al., 1999; Delen et al., 2006).

To overcome the limitations associated with traditional statistical models, previous researchers have proposed non-parametric methods and artificial intelligence models for crash injury severity analysis. These models include the classification and regression tree (CART) model (Sohn and Shin, 2001; Karlaftis and Golias, 2002; Chang and Wang, 2006), the Bayesian network model (Simoncic, 2004; de Ona et al., 2011), and the artificial neural network models (Abdelwahab and Abdel-Aty, 2002; Delen et al., 2006; Xie et al., 2007), etc. As compared to traditional parametric models such

* Corresponding author. Tel.: +86 025 83791816.
*E-mail addresses:* lizhibin@seu.edu.cn (Z. Li), pan_liu@hotmail.com (P. Liu), wangwei@seu.edu.cn (W. Wang), iamxcc@163.com (C. Xu).
[1] Tel.: +86 13952097374.
[2] Tel.: +86 13905170160.
[3] Tel.: +86 13801580045.

0001-4575/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2011.08.016

as the OP model, non-parametric models, including the artificial intelligence models do not need any pre-defined underlying relationships between dependent variables and explanatory variables. Some previous studies also showed that non-parametric models usually produced better statistical fit than traditional parametric models (Fish and Blodgett, 2003; Xie et al., 2007; de Ona et al., 2011).

Support vector machine (SVM) model is a relatively new modeling technique which was proposed to solve the classification and regression problems (Kecman, 2005). In recent years, the SVM model has been widely used in transportation studies including traffic flow prediction (Cheu et al., 2006; Zhang and Xie, 2007), incident detection (Yuan and Cheu, 2003; Chen et al., 2009a), travel mode choice modeling (Zhang and Xie, 2008), and crash frequency prediction (Li et al., 2008), etc. It was found that the SVM model has great capability for dealing with classification problems. Thus, it also can be used for modeling crash injury severity data which are categorical in nature. The major limitation associated with the SVM model lies in the fact that the model generally works like a black-box which cannot be directly used to identify the relationships between crash injury severity and various explanatory variables.

In recent years, sensitivity analysis method has been proposed to minimize the black-box problems associated with artificial intelligence models and to identify the effects of explanatory variables (Olden et al., 2004). The use of the sensitivity analysis method has greatly expanded the potential of using SVM model in traffic safety analysis (Li et al., 2008). It was found that SVM models produced better or at least comparable estimation results as compared to traditional statistical models (Li et al., 2008; Zhang and Xie, 2008). However, until recently, the SVM models have never been used for crash injury severity analysis.

The primary objective of this study is to investigate the possibility of using SVM model for crash injury severity analysis. More specifically, the study presented in this paper focused on answering the following two questions: (1) if SVM model provides better estimation results regarding crash injury severities as compared to traditional OP model; and (2) if SVM model can be used to estimate the effects of contributing factors on crash injury severity.

## 2. Data and methods

### 2.1. Data sources

Crash data used for injury severity analysis were obtained from 326 freeway segments in the State of Florida, United States. The freeway segment defined in this study is a section of freeway which contains a deceleration lane and an exit. The freeway segment consists of two influence areas, including (1) an influence area located within 457 m (1500 ft) upstream of the painted nose and (2) an influence area located within 305 m (1000 ft) downstream of the pained nose. Thus, the freeway segment defined in this study has a consistent length of 762 m (2500 ft).

The freeway exits were classified into four different types based on the number and arrangement of lanes used for traffic to exit freeways. They were defined as type 1, type 2, type 3, and type 4 exit ramps, respectively. Type 1 exit ramp is a single lane exit ramp with a tapered design. Type 2 exit ramp is a single lane exit ramp with the outer lane of freeway becomes a drop lane at the exit gore. Type 3 exit ramp is a two-lane exit with an optional lane. Type 4 exit ramp is a two-lane exit with the outer lane of the freeway dropped at the exit gore. The definition of each type of exit ramp is illustrated in Fig. 1.

A three-year time frame, from 2004 to 2006, was defined for obtaining crash data at selected sites. In 2003, the FDOT renamed all of the freeway exit ramps for the whole state and, accordingly, the crash database updated the exit ramp numbers for the entire database. The research team found that the crash data for freeway exit ramps before 2004 has some missing information and, sometimes cannot be matched with the data after 2004. Due to this reason, we only selected crash data after 2004 for further analysis.

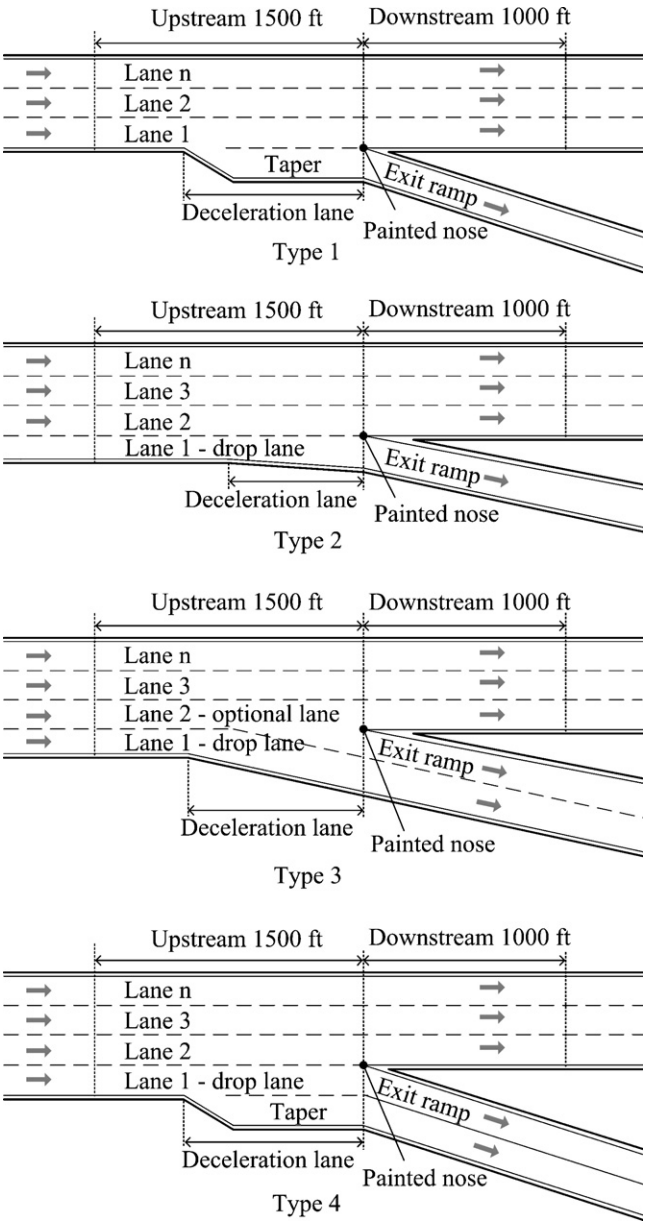


**Fig. 1.** Illustrations of four types of exit ramps on freeways.

A total of 5538 crashes reported at selected freeway segments were used for injury severity analysis in this study. Descriptions of injury severity of crashes were given in Table 1. The crash injury

**Table 1**
Injury severity levels represented in dataset.

| Injury severity level | Description | Frequency | Percent |
|---|---|---|---|
| Level 1 | No injury | 2902 | 52.5 |
| Level 2 | Possible/invisible injury | 1463 | 26.4 |
| Level 3 | No-capacitating injury | 837 | 15.1 |
| Level 4 | Incapacitating injury | 285 | 5.1 |
| Level 5 | Fatal injury | 51 | 0.9 |
| Overall | – | 5538 | 100.0 |

**Table 2**
Descriptions of selected variables for analysis.

| Variable | Description | Frequency | Percent |
|---|---|---|---|
| Ramp type | Type 1 Exit Ramp | 2387 | 43.1 |
| | Type 2 Exit Ramp | 1848 | 33.4 |
| | Type 3 Exit Ramp | 948 | 17.1 |
| | Type 4 Exit Ramp | 355 | 6.4 |
| Main lanes | 2 Lanes on mainline | 535 | 9.7 |
| | 3 Lanes on mainline | 1408 | 25.4 |
| | 4 Lanes on mainline | 1298 | 23.4 |
| | 5 Lanes on mainline | 1533 | 27.7 |
| | 6 Lanes on mainline | 765 | 13.8 |
| Ramp lanes | 1 Lane on exit ramp | 4180 | 75.5 |
| | 2 Lanes on exit ramp | 1308 | 23.6 |
| | 3 Lanes on exit ramp | 51 | 0.9 |
| DeLength | Length of deceleration lanes (mile) | Continuous | |
| RaLength | Length of entire exit ramps (mile) | Continuous | |
| SurfacType | 1 Blacktop surface | 4390 | 79.3 |
| | 0 others | 1148 | 20.7 |
| ShoulderType | 1 Paved shoulder | 3840 | 69.3 |
| | 0 no Paved shoulder | 1698 | 30.7 |
| ShoulderWidth | Right shoulder width (ft) | Continuous | |
| MainSpeed | Post speed limit on mainline (mph) | Continuous | |
| SpeedDiff | Difference of speed limit between mainline and exit ramps (mph) | Continuous | |
| Light | 1 Daylight (including dusk and dawn) | 3785 | 68.3 |
| | 0 No daylight | 1753 | 31.7 |
| Weather | 1 Clear weather condition | 3647 | 65.8 |
| | 0 Others | 1891 | 34.2 |
| Surface | 1 Wet surface condition | 1148 | 20.7 |
| | 0 Dry surface condition | 4390 | 79.3 |
| LandType | 1 Business surroundings | 3185 | 57.5 |
| | 0 Residential surroundings | 2353 | 42.5 |
| MainADT | Mainline ADT per year in thousand | Continuous | |
| RampADT | Exit ramp ADT per year in thousand | Continuous | |
| AlcDrug | 1 Alcohol/drug involved | 223 | 0.4 |
| | 0 No alcohol/drug involved | 5315 | 99.6 |
| Crash type | Rear end crash | 2347 | 42.4 |
| | Sideswipe crash | 760 | 13.7 |
| | Angle crash | 450 | 8.1 |
| | Others | 1981 | 35.8 |

severity includes five ordered levels ranking from level 1 to level 5. Level 1 represents no-injury crashes which account for more than half of all crashes (52.5%). Level 2 represents possible/invisible injury crashes which account for 26.4% of total crashes. Level 3 and level 4 denote no-capacitating injury crashes and incapacitating injury crashes, which account for 15.1% and 5.1% of all crashes, respectively. Level 5 represents fatal crashes with the smallest proportion of 0.9%. The database used for injury severity analysis also includes various explanatory variables for each crash, such as the roadway geometric design characteristics, environment/traffic conditions and the information for crashes. Descriptive statistics of explanatory variables are given in Table 2. Since the focus of this paper is not on the safety performance of freeway exit ramps, detailed data collection procedure was not discussed. For more details regarding the data sources, please refer to our previous study (Chen et al., 2009b).

## 2.2. Methodology

The models used for crash injury severity analysis are briefly discussed in this section. The SVM model treats the crash injury severity modeling as a classification problem, i.e., the crashes were classified into different categories according to their severity level. The SVM model can be used to predict the injury severity given the fact that a crash has occurred. For comparison purposes, we also developed a traditional OP model based on the same dataset. As compared to the SVM model, the OP model treats crash injury severity as an ordinal variable and estimates the relationships between injury severity and explanatory factors.

### 2.2.1. SVM model

The SVM model is based on the statistical learning theory and the structural risk minimization principle. To introduce the classification function of SVM model, a simple two-category separable classification problem is illustrated in Fig. 2(a). The goal of using SVM model is to separate the two classes of samples. As shown in the figure, a SVM model can map the input vector, $\mathbf{X}$, into a high dimensional feature space. By choosing a non-linear mapping a priori, the SVM model constructs an optimal separating hyperplane in this higher dimensional space to separate the outcome into several groups, while maximizing the margin between the linear decision boundaries (shown as dashed line in Fig. 2(a)). Fig. 2(b) shows the structure of classification in SVM model.

When specifying the SVM model, the data is separated into a training set and a testing set. The SVM model produces a learning model based on the training set and then makes predictions of the testing set. The training input is defined as vectors $\boldsymbol{x}_i \in R^n$ for $i = 1,2,\ldots,N$, which represent the full set of crash-related variables, and the training output is defined as $\boldsymbol{y}_i \in R^n$, which represents the injury severity of crashes. The hyperplane for separating outcomes can be written as the set of points $\mathbf{X}$ satisfying:

$$\mathbf{W} \cdot \mathbf{X} - b = 0 \tag{1}$$

where "·" denotes the dot product and the vector $\mathbf{W}$ is a normal vector: it is perpendicular to the hyperplane. For a two-category classification task, given a training set of instance-label pairs ($x_i$,

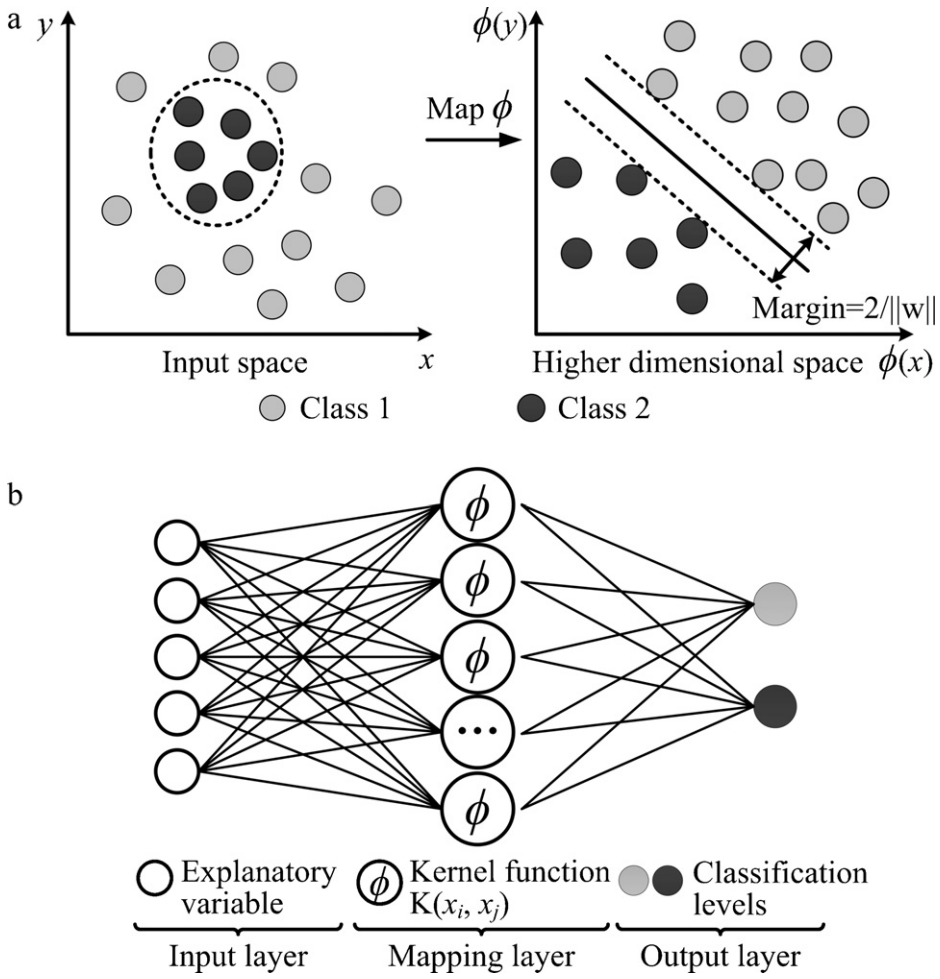


**Fig. 2.** Graphical representation of classification in SVM model.

$y_i$), the SVM model requires to solve the following optimization problem (Cortes and Vapnik, 1995):

$$\begin{aligned} &\min_{w,b,\xi} \tfrac{1}{2} w^T w + C\sum_{i=1}^{N}\xi_i \\ &\text{Subject to} \quad y_i(w^T\phi(x_i)+b) \geq 1-\xi_i, \\ &\qquad\qquad \xi_i \geq 0. \end{aligned} \tag{2}$$

where $\xi$ are slack variables measuring the misclassification errors, and $C$ is the penalty factor to errors introducing additional capacity control within the classifier. The uncertain part of the above approach is that the coefficient $C$ has to be determined. This constraint along with the objective of minimizing function can be solved using Lagrange multipliers. One has then to solve the following problem:

$$\min \max \left\{ \frac{1}{2} w^T w + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(w^T\phi(x_i)+b)-1+\xi_i] - \sum_{i=1}^{n}\beta_i\xi_i \right\} \tag{3}$$

where $\alpha_i$, $\beta_i > 0$ are Lagrange multipliers. Here training vectors $x_i$ are mapped into a higher (maybe infinite) dimensional space by the function $\phi(x_i)$. Furthermore, $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$ is called the kernel function. Though several kernels have been proposed by researchers, a most commonly used radial basis function (RBF) was used for crash injury severity analysis in this study, putting aside other types of functions in future test. The radial basis function is defined as:

$$K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2), \gamma > 0 \tag{4}$$

where $\gamma$ is the kernel parameter. With the radial basis function as the kernel function, the SVM model has two parameters ($C$, $\gamma$) that need to be determined. There is always a globally optimal solution to **W** and $b$ with the input of parameters ($C$, $\gamma$).

The SVM model demonstrated so far is for the two-category classification problem. This model can be easily extended for dealing with multi-category classification tasks. The crash injury severity in this study includes five category levels. The prevailing one-versus-one approach is adopted to solve multi-level injury severity classification problem. This approach conducted classification by a max-wins voting strategy, in which every classifier assigns the crash severity to one of the two severity levels (one vote). The class with the most votes determines the severity level of the crash. For more details regarding the multi-category classification of SVM model, please refer to Lingras and Butz (2007).

Fig. 3 shows the flow chart of the SVM model used for crash injury severity analysis in this study. Each crash in the training dataset has a label which indicates the severity level (fatal, injury etc.) and its paired individual crash information (crash type, speed limit etc.). The SVM model learned the relationships between crash injury severity and explanatory variables based on the training crash data. The optimal values of parameter $C$ and $\gamma$ were identified

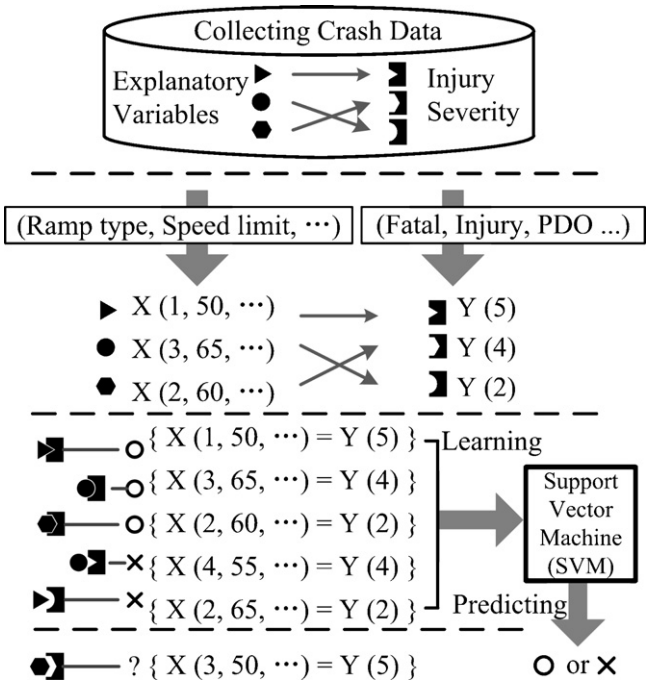


**Fig. 3.** Graphical representation of SVM model for crash injury severity analysis.

during the learning process to maximize the model performance. Then given crash information in the testing crash dataset, the SVM model can make prediction on the severity level of each crash. By comparing the predicted and observed severities of crashes, the accuracy which indicates the proportion of correctly classified crashes can be calculated. Accuracy gives information on the SVM classifier's general performance.

As mentioned before, the SVM model has been long criticized for performing as a black-box which cannot be directly used to identify the relationship between outcomes and input variables. Recently, artificial intelligence researchers have proposed mathematical methods to infer the estimates of artificial intelligence models (Fish and Blodgett, 2003; Olden et al., 2004). Sensitivity analysis is one of the methods used to evaluate the relationship between inputs and outputs (Delen et al., 2006; Xie et al., 2007; Li et al., 2008). This study used sensitivity analysis in SVM model to infer the impacts of contributing factors on crash injury severity.

#### 2.2.2. OP model

The OP model has been widely used for fitting the data structure of an ordinal response. Crash injury severity can be defined as typical ordinal variables which are scaled into five levels according to the Florida crash database used in this study: 1-no injury; 2-possible/invisible injury; 3-non-incapacitating injury; 4-incapacitating injury; and 5-fatal injury. Assuming that $\mathbf{Y}$ represents the injury severity level, then a latent variable $\mathbf{Y}^*$ is given as:

$$\mathbf{Y}^* = \mathbf{X}\beta + \boldsymbol{\varepsilon} \tag{5}$$

where $\mathbf{X}$ is the vector containing the full set values of crash-related variables, $\boldsymbol{\beta}$ is the vector of coefficients associated with the explanatory variables, and $\boldsymbol{\varepsilon}$ is a random error term following a standard normal distribution. The value of the dependent variable $\mathbf{Y}$ is then given as (C.f. Washington et al., 2003):

$$Y = \begin{cases} 1 & \text{if } Y^* < \tau_1 \\ j & \text{if } \tau_{j-1} < Y^* < \tau_j \\ J & \text{if } \tau_{J-1} < Y^* \end{cases} \tag{6}$$

where $J$ is the number of injury severity levels (in this case, $J=5$), and $\tau_j$ is the threshold parameter (cut-off points) to be estimated for each severity level. The probability for $\mathbf{Y}$ taking a particular value $j$ is given by:

$$\begin{aligned} P(Y=1) &= \Phi(\tau_1 - \mathbf{X}\beta) \\ P(Y=j) &= \Phi(\tau_j - \mathbf{X}\beta) - \Phi(\tau_{j-1} - \mathbf{X}\beta) \\ P(Y=J) &= 1 - \Phi(\tau_{J-1} - \mathbf{X}\beta) \end{aligned} \tag{7}$$

where $P(Y=j)$ is the probability of response variable taking a specific severity level $j$, $\Phi(\cdot)$ is the standard normal cumulative distribution function, and the threshold parameter $\tau_j$ satisfies the restriction $\tau_1 < \tau_2 < \cdots < \tau_{J-1}$. For the classical OP model, the values of $\boldsymbol{\beta}$ and $\mathbf{Y}^*$ can be determined by the Maximum Likelihood estimate method. The likelihood function can be given as:

$$\begin{aligned} L &= L(Y|\tau_1, \tau_2, \tau_3, \tau_4, \beta_0, \beta_1, \cdots, \beta_n) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{J}\{\Phi(\tau_j - \beta' X_i) - \Phi(\tau_{j-1} - \beta' X_i)\}^{Y,i+5} \end{aligned} \tag{8}$$

where $N$ is the number of explanatory variables. The OP model can be directly used to evaluate the impacts of explanatory variables on probability of each injury severity by calculating the marginal effects of contributing factors. For continuous explanatory variables, the marginal effect of a variable $i$ for injury severity $j$, $\Delta P(Y=j|x_i)$, is given by:

$$\Delta P(Y=j|x_i) = P(Y=j)/\partial x_i = [\Phi(\tau_{j-1} - \boldsymbol{\beta}\mathrm{X}) - \Phi(\tau_j - \boldsymbol{\beta}\mathrm{X})]\beta_i \tag{9}$$

For binary (dummy) variables, the marginal effect of a variable $i$ for injury severity $j$ is given by:

$$\Delta(Y=j|x_i) = P(Y=j|x_i=1) - P(Y=j|x_i=0) \tag{10}$$

## 3. Data analysis results

### 3.1. Model specification

#### 3.1.1. SVM model

The LIBSVM tool developed by Chang and Lin (2007) in MATLAB (The MathWorks, Inc. 2009) was used for specifying the SVM model in this study. The LIBSVM tool provides a grid searching algorithm for determining the parameters $(C, \gamma)$ of the SVM model. Original crash dataset was randomly separated into a training set and a testing set with a ratio of 4:1. In order to reduce the bias associated with random separation of crash dataset, ten experiments with different training and testing sets were conducted for injury severity analysis. The mean classification accuracy of ten estimates was considered the model performance, as shown in Table 3.

The overall classification accuracy based on the training set is 79.6%, indicating the fact that 79.6 percent of injury severities associated with individual crashes was correctly identified. The overall correct classification rate for the testing set is 48.8%. In general, the calibrated SVM model performs worse on the testing set than on the training set. It was also found that the SVM model performs better on injury severities with larger proportions. In other words, the estimation accuracy for no-injury crashes and possible/invisible injury crashes are generally higher than that for fatal and incapacitating injury crashes. This is a typical multi-class classification problem which has been commonly observed in classification technologies such as classification tree, artificial neural network, and SVM models, etc. (Chang and Wang, 2006; Delen et al., 2006; Zhang and Xie, 2008). The problem is that the fatal injuries and incapacitating injures take only small proportions in the crash database (5.1% and 0.9%, respectively). The SVM model simply ignored the

**Table 3**
Classification accuracy of SVM and OP models.

| Injury severity | Training (%) | | Testing (%) | |
|---|---|---|---|---|
| | Mean | S.D. | Mean | S.D. |
| SVM | | | | |
| No injury | 93.4 | 0.514 | 77.0 | 4.618 |
| Possible/invisible injury | 68.1 | 2.836 | 25.4 | 5.573 |
| No-capacitating injury | 61.2 | 1.984 | 10.1 | 3.107 |
| Incapacitating injury | 55.3 | 3.359 | 2.3 | 1.642 |
| Fatal injury | 65.4 | 5.339 | 1.7 | 3.408 |
| *Overall* | 79.6 | 1.537 | 48.8 | 4.475 |
| No injury | 86.4 | 1.262 | 61.3 | 1.239 |
| Injury/fatal | 81.0 | 1.311 | 53.6 | 0.787 |
| *Overall* | 83.8 | 0.183 | 57.6 | 0.680 |
| OP | | | | |
| No injury | 58.1 | 0.697 | 56.8 | 2.680 |
| Possible/invisible injury | 54.1 | 1.867 | 52.6 | 3.086 |
| No-capacitating injury | 0.1 | 0.127 | 0.1 | 0.296 |
| Incapacitating injury | 0.0 | 0.000 | 0.0 | 0.000 |
| Fatal injury | 0.0 | 0.000 | 0.0 | 0.000 |
| *Overall* | 44.7 | 0.359 | 44.0 | 1.961 |

information from small categories in order to improve the overall classification accuracy.

One popular approach for dealing with the multi-class classification problem and producing better predictions is to reduce the multi-class classification task into a two-class classification task (Tax and Duin, 2002; Delen et al., 2006). In this study, the five injury severities were then aggregated into two severity levels, including a no-injury level and an injury/fatal level. The classification results of the two-level SVM model were also given in Table 3. The results showed that the prediction accuracy of the SVM model was significantly improved.

#### 3.1.2. OP model

The OP model was fitted using the STATA software package. The estimation results of OP model are given in Table 4. Based on the OP model, the variables which significantly affect the injury severity of crashes include: the number of lanes on freeway mainline, the type of land use in surrounding area, the length of the entire exit ramp, the shoulder width of freeway mainline, freeway pavement surface conditions, lighting conditions, weather conditions, alcohol/drug involvement, and rear-end and sideswipe collision types. The estimated coefficients for each contributing factor and the threshold value for each injury severity category are also given in Table 4.

### 3.2. Comparison of prediction performance

One of the tasks of this study is to identify if the SVM model has better prediction performance than OP model. To make the comparison more reasonable, the OP model was also fitted using randomly separated fitting and testing datasets. The fitting data was using to fit the model while the testing dataset was used to evaluate the prediction performance of the model. The parameter accuracy which is the percent of correct predictions was used for comparing the prediction performance of models. The parameter accuracy can be calculated using Eq. (11).

$$\text{Accuracy}: \frac{\text{the number of correct predictions}}{N} \tag{11}$$

where $N$ is the number of observations in the testing dataset.

The prediction accuracy of the OP model was also given in Table 3. It was found that the SVM model produced better prediction performance for crash injury severity than did the OP model. The percent of correct prediction for the SVM model was found to be 48.8%, which was higher than that produced by the OP model (44.0%).

The research team also tested the prediction performance of OP model on small proportion injury severities such as the fatal and incapacitating injury. It was found that the multi-class classification problem also exits in the OP model. In fact, none of the crashes in the testing set was predicted to be a fatal or incapacitating injury crash by the OP model. The results show that even though the SVM model may suffer from the multi-class classification problem, it still provides better prediction results for small proportion injury severities than the OP model does.

### 3.3. Impacts of explanatory variables

Sensitivity analysis was conducted in this study to explore the relationship between crash injury severity and various explanatory variables using the SVM model. The learning process of the SVM model was disabled to ensure that the model was not affected by the change in the testing dataset. Each input variable of the SVM model was perturbed by a user-defined amount with other variables remain unchanged. The estimation result before and after the perturbation of each input variable was recorded. The impacts of each input variable on crash injury severity were then estimated as the percent change of each severity level by one unit change of each input variable. Note that we only focused on the impacts of the variables which were found to be statistically significant in the OP model.

**Table 4**
Estimation results of OP model.

| | | | | | | |
|---|---|---|---|---|---|---|
| Number of crashes = 5538 | | LR $chi^2$ (10) = 240.20 | | | | |
| | | Prob > $chi^2$ = 0.0000 | | | | |
| Log likelihood = −6369.2043 | | Pseudo $R^2$ = 0.0187 | | | | |
| Variables | Coef. | Std. Err. | $z$ | $P>\|z\|$ | [95% Conf. interval] | |
| MainLanes | 0.0513 | 0.0131 | 3.92 | 0.000 | 0.0257 | 0.0770 |
| RaLength | 0.1365 | 0.0575 | 2.38 | 0.018 | 0.0238 | 0.2492 |
| LandType | 0.1239 | 0.0325 | 3.81 | 0.000 | 0.0602 | 0.1875 |
| ShoulderWidth | 0.2446 | 0.0472 | 5.18 | 0.000 | 0.1521 | 0.3370 |
| Surface | −0.1007 | 0.0506 | −1.99 | 0.047 | −0.1999 | −0.0015 |
| Light | −0.0856 | 0.0339 | −2.52 | 0.012 | −0.1520 | −0.0191 |
| Weather | −0.0820 | 0.0401 | −2.05 | 0.041 | −0.1605 | −0.0035 |
| AlcDrug | 0.3752 | 0.0764 | 4.91 | 0.000 | 0.2254 | 0.5249 |
| RearEnd | −0.0710 | 0.0333 | −2.13 | 0.033 | −0.1362 | −0.0058 |
| Sideswipe | −0.6075 | 0.0517 | −11.74 | 0.000 | −0.7089 | −0.5061 |
| /cut1 | 0.9288 | 0.1660 | | | 0.6034 | 1.2542 |
| /cut2 | 1.6894 | 0.1667 | | | 1.3627 | 2.0160 |
| /cut3 | 2.4559 | 0.1681 | | | 2.1265 | 2.7853 |
| /cut4 | 3.2863 | 0.1744 | | | 2.9445 | 3.6281 |

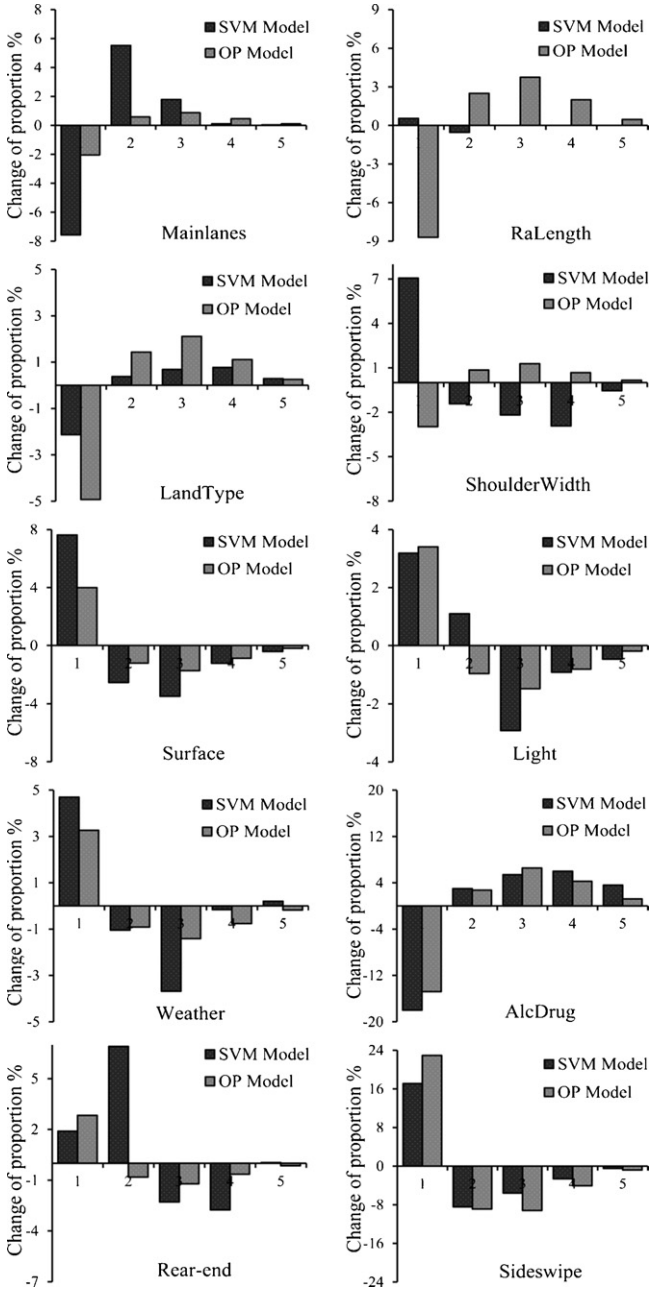


**Fig. 4.** Comparison of SVM and OP models regarding the impacts of variables on crash injury severity.

For comparison purpose, the marginal effects of contributing factors on each injury severity level were estimated using the OP model. The marginal effect of a variable in the OP model represents the change of the probability of each severity level by one unit increase in the input variable. The sensitivity analysis results and the marginal effects of each variable are compared in Fig. 4.

As shown in Fig. 4, as compared to the OP model, the SVM model produced comparable results regarding the impacts of variables on crash injury severity. Of the ten variables considered, the SVM model and the OP model produced similar results for eight explanatory variables. Those variables are: the number of lanes on freeway mainline, the type of land use in surrounding area, freeway pavement surface conditions, lighting conditions, weather conditions, alcohol/drug involvement, and rear-end and sideswipe collision types.

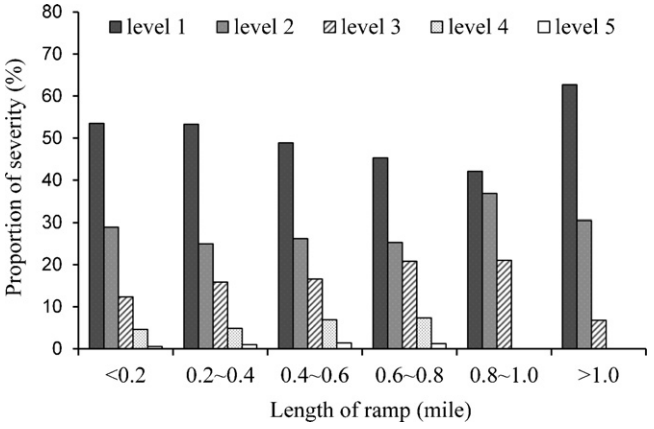


**Fig. 5.** Relationship between injury severity and ramp length.

The SVM model and the OP model produced inconsistent results for two variables, including the shoulder width of freeway mainline and the length of the entire exit ramp. Based on the OP model, the increase in the length of the entire exit ramp will significantly increase crash injury severity at freeway diverge areas. However, the SVM model shows that the impacts of the length of exit ramp are minor. Theoretically, the length of the entire exit ramp should not have direct impacts on injury severity at freeway diverge areas since the crashes on exit ramps were not considered. To further evaluate the impacts of ramp length on injury severity, distributions of crash injury severity against ramp length are plotted in Fig. 5. As shown in Fig. 5, the length of the exit ramp generally has minor impacts on the distribution of injury severities, indicating the fact that the result provided by the SVM model is more reasonable.

For the shoulder width of freeway mainlines, the SVM and the OP model provided completely opposite results. The OP model shows that the increase in shoulder width will increase crash severity, while the SVM model produces the opposite conclusion. Again, the estimate of the OP model is quite counter-intuitive. To identify the truth, the relationship between injury severity and shoulder width based on the original dataset was plotted in Fig. 6. The solid line in Fig. 6 shows that the mean injury severity value generally increases with the increase of shoulder width. It looks like the OP model provided better estimates this time. However, by examining the dataset, it was found that only 1 site (12 crash samples) had a 1 ft wide shoulder, and the minimum shoulder width of the other sites was 6 ft. It we consider the 1 site as outliers, the

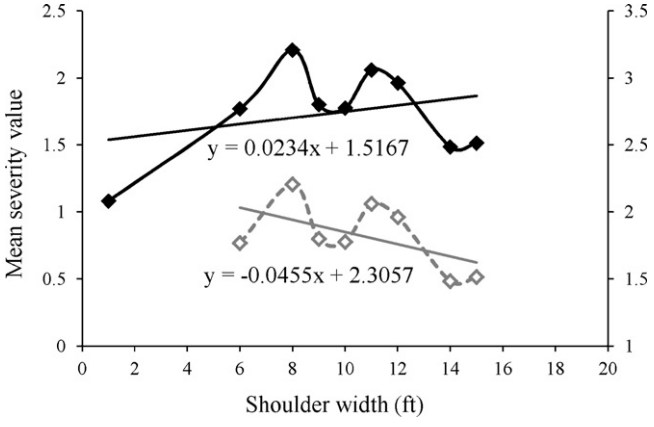


**Fig. 6.** Relationship between injury severity and shoulder width.

relationship between the shoulder width and crash injury for other 325 sites is illustrated in Fig. 6 as the dashed line. It was found that the increase of shoulder width would decrease the injury severity, and the result is consistent with the estimate of the SVM model. During model specification, the SVM model ignored the information from the 12 samples to get a better overall prediction performance.

## 4. Conclusions and discussion

The study presented in this paper investigated the possibility of using SVM models for crash injury severity analysis. Based on crash data collected at 326 freeway diverge areas, a SVM model was developed for predicting the injury severity associated with individual crashes. An OP model was also developed using the same dataset. The research team compared the prediction performance of the SVM model and the OP model. It was found that the SVM model produced better prediction performance for crash injury severity than did the OP model. The percent of correct prediction for the SVM model was found to be 48.8%, which was higher than that produced by the OP model (44.0%). Even though the SVM model may suffer from the multi-class classification problem, it still provides better prediction results for small proportion injury severities than the OP model does.

The major conclusion of this study is that the SVM model can be used for crash injury severity analysis. Using the sensitivity analysis, the SVM model can also be used to evaluate the impacts of explanatory variables on crash injury severity. The sensitivity analysis results show that the SVM produced comparable results regarding the impacts of variables on crash injury severity as compared to the OP model. For several variables such as the length the exit ramp and the shoulder width of the freeway mainline, the results of the SVM model is more reasonable than those of the OP model.

This study does have several limitations. For example, the performance of the SVM model highly depends on the learning procedure which contains functional mapping and parameter selection. In this study, we only used the basic radial basis function (RBF). Other kernel functions can be tested to improve the model performance. Other parameter searching algorithms, such as the Genetic Algorithm, may also be used to improve the performance of the SVM model. In addition, the model partly demonstrated the benefits of using the SVM model in crash injury severity analysis over traditional OP models. However, the research team did not compare the SVM model to other statistical models such as the multinomial logit model and logistic regression model. The study was based on crash data collected during a three-year time frame. It was assumed that confounding factors which were not included in injury severity models during the three-year observation period remained constant. Even though estimating crash severity at freeway exits is not the primary focus of this study, the change of confounding factors over the study period may affect the estimation results of injury severity models. The authors recommend that future studies may focus on these issues.

## Acknowledgments

This research was jointly sponsored by China's National Science and Technology Plan of Action for Traffic Safety (Project #: 2009BAG13A07-5), China's National Science Foundation (Project #: 50908050), as well as the key project of the National Natural Science Foundation of China (No. 34 50738001).

## References

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34 (5), 597–603.
Abdelwahab, H.T., Abdel-Aty, M.A., 2002. Artificial neural networks and logit models for traffic safety analysis of toll plazas. Transportation Research Record: Journal of the Transportation Research Board 1784, 115–125.
Abdel-Aty, M.A., Abdelwahab, H.T., 2004. Predicting injury severity levels in traffic crashes: A modeling comparison. Journal of Transportation Engineering-Asce 130 (2), 204–210.
Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis and Prevention 34 (6), 729–741.
Chang, C.C., Lin, C.J., 2007. LIBSVM: A Library for Support Vector Machines, accessed on July, 2010 www.csie.ntu.edu.tw/cjlin/libsvm.
Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38 (5), 1019–1027.
Chen, H.Y., Liu, P., Lu, J.J., Behzadi, B., 2009a. Evaluating the safety impacts of the number and arrangement of lanes on freeway exit ramps. Accident Analysis and Prevention 41 (3), 543–551.
Chen, S.Y., Wang, W., Henk, J.Z., 2009b. Construct support vector machine ensemble to detect traffic incident. Expert Systems with Applications 36 (8), 10976–10986.
Cheu, R.L., Xu, J., Kek, A.G.H., Lim, W.P., Chen, W.L., 2006. Forecasting shared-use vehicle trips with neural networks and support vector machines. Transportation Research Record: Journal of the Transportation Research Board 1968, 40–46.
Cortes, C., Vapnik, V., 1995. Support-vector network. Machine Learning 20, 273–297.
de Ona, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury severity on spanish rural highways using bayesian networks. Accident Analysis and Prevention 43 (1), 402–411.
Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis and Prevention 38 (3), 434–444.
Duncan, C.S., Khattak, A.J., Council, F.M., 1998. Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions. Transportation Research Record: Journal of the Transportation Research Board 1635, 63–71.
Fish, K.E., Blodgett, J.G., 2003. A visual method for determining variable importance in an artificial neural network model: an empirical benchmark study. Journal of Targeting, Measurement and Analysis for Marketing 11, 244–254.
Karlaftis, M.G., Golias, I., 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accident Analysis and Prevention 34 (3), 357–365.
Kecman, V., 2005. Support vector machines—an introduction. In: Wang, L. (Ed.), Support Vector Machines: Theory and Applications. Springer-Verlag, Berlin, Heidelberg, New York, pp. 1–48.
Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F., 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accident Analysis and Prevention 37 (5), 910–921.
Kockelman, K.M., Kweon, Y.J., 2002. Driver injury severity: an application of ordered probit models. Accident Analysis and Prevention 34 (3), 313–321.
Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle-pedestrian crashes at intersections in florida. Accident Analysis and Prevention 37 (4), 775–786.
Li, X.G., Lord, D., Zhang, Y.L., Me, Y.C., 2008. Predicting motor vehicle crashes using support vector machine models. Accident Analysis and Prevention 40 (4), 1611–1618.
Lingras, P., Butz, C., 2007. Rough set based 1-v-1 and 1-v-r approaches to support vector machine multi-classification. Information Sciences 177 (18), 3782–3798.
Mussone, L., Ferrari, A., Oneta, M., 1999. An analysis of urban collisions using an artificial intelligence model. Accident Analysis and Prevention 31 (6), 705–718.
Odonnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accident Analysis and Prevention 28 (6), 739–753.
Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling 178 (3-4), 389–397.
Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Analysis and Prevention 39 (5), 955–963.
Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. Journal of Safety Research 27 (3), 183–194.
Siddiqui, N.A., Chu, X.H., Guttenplan, M., 2006. Crossing locations, light conditions, and pedestrian injury severity. In: The 85th Annual Meeting of the Transportation Research Board, Washington, D.C.
Simoncic, M., 2004. A Bayesian network model of two-car accidents. Journal of Transportation and Statistics 7 (2/3), 13–25.
Sohn, S.Y., Shin, H., 2001. Pattern recognition for road traffic accident severity in Korea. Ergonomics 44 (1), 107–117.
Tax, D.M.J., Duin, R.P.W., 2002. Using two-class classifiers for multiclass classification. In: Proceedings of the 16th International Conference on Pattern Recognition, pp. 124–127.
Wang, Z.Y., Cao, B., Deng, W., Zhang, Z., Lu, J.J., Chen, H.Y., 2011. Safety evaluation of truck-related crashes at freeway diverge areas. In: The 90th Annual Meeting of the Transportation Research Board, Washington, D.C.

Xie, Y.C., Lord, D., Zhang, Y.L., 2007. Predicting motor vehicle collisions using bayesian neural network models: an empirical analysis. Accident Analysis and Prevention 39 (5), 922–933.
Xie, Y.C., Zhang, Y.L., Liang, F.M., 2009. Crash injury severity analysis using bayesian ordered probit models. Journal of Transportation Engineering 135 (1), 18–25.
Yau, K.K.W., Lo, H.P., Fung, S.H.H., 2006. Multiple-vehicle traffic accidents in hong kong. Accident Analysis and Prevention 38 (6), 1157–1161.
Yuan, F., Cheu, R.L., 2003. Incident detection using support vector machines. Transportation Research Part C: Emerging Technologies 11 (3-4), 309–328.
Zajac, S.S., Ivan, J.N., 2003. Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural connecticut. Accident Analysis and Prevention 35 (3), 369–379.
Zhang, Y.L., Xie, Y.C., 2007. Forecasting of short-term freeway volume with v-support vector machines. Transportation Research Record: Journal of the Transportation Research Board 2024, 92–99.
Zhang, Y.L., Xie, Y.C., 2008. Forecasting of short-term freeway volume with v-support vector machines. Transportation Research Record: Journal of the Transportation Research Board 2076, 141–150.