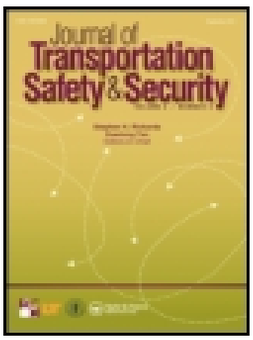

Journal of Transportation Safety & Security

ISSN: 1943-9962 (Print) 1943-9970 (Online) Journal homepage: https://www.tandfonline.com/loi/utss20

A novel ensemble learning method for crash prediction using road geometric alignments and traffic data

Peijie Wu, Xianghai Meng & Li Song

To cite this article: Peijie Wu, Xianghai Meng & Li Song (2019): A novel ensemble learning method for crash prediction using road geometric alignments and traffic data, Journal of Transportation Safety & Security, DOI: 10.1080/19439962.2019.1579288

To link to this article: https://doi.org/10.1080/19439962.2019.1579288

Published online: 06 Jun 2019.

JOURNAL OF TRANSPORTATION SAFETY & SECURITY
https://doi.org/10.1080/19439962.2019.1579288

A novel ensemble learning method for crash prediction using road geometric alignments and traffic data

Peijie Wu, Xianghai Meng, and Li Song

School of Transportation Science and Engineering, Harbin Institute of Technology, Heilongjiang, China

ABSTRACT

As an important part of traffic safety analysis, crash prediction models using road geometric alignments and traffic data (CPM-GAs) have been regarded as the most classic way and can be used in stages of road safety evaluation and road operating and management. To improve the predictive performance of tradition CPM-GAs and avoid the overfitting problem of machine learning algorithms, a framework of CPM-GA based on ensemble learning theory and a new ensemble rule for connecting traditional models and machine learning models were proposed in this study. Results of the ensemble learning CPM-GA show that (1) classification and regression tree (CART) is recommended for important variable selection procedure before applying support vector machine (SVM), (2) machine learning models outperformed traditional models significantly in aspects of model fitting and prediction accuracy but are unstable in the sensitivity tests, (3) the new proposed ensemble method of traditional model and machine learning model can effectively improve the accuracy of traditional CPM-GAs by 10%–16% and reduce the variance of machine learning CPM-GAs by 12%–36% simultaneously. Finally, the ensemble method presented in this article may shed light on more research of crash prediction models based on ensemble learning theory.

KEYWORDS

traffic safety; crash prediction model; ensemble learning; support vector machine; interactive highway safety design model; road geometric design

1. Introduction

In the advent of rapid economic growth and huge travel demand of China, national highway network plan (NHNP) was officially proposed in 2013 (Traffic Management Bureau of Ministry of Public Security, 2013). However, when the large-scale freeway network offers convenience and efficiency, it is also associated with increased crash frequency. There were 8,693 crashes occurring on freeways in China, of which 4.38% were reported to police (Traffic Management Bureau of Ministry of Public Security, 2014). Therefore, improving traffic-safety level of freeways in

CONTACT Xianghai Meng mengxianghai100@126.com No.73. Huanghe Road, Nangang District, School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin, Heilongjiang, 150090, China.

China is becoming the first priority for Chinese freeway management department. To accomplish this goal, traffic safety improvements need to be conducted in high-risk freeway segments. As the core of traffic safety analysis, accurate and effective crash prediction models (CPMs) are in need. Crash prediction models using road geometric alignments and traffic data (CPM-GAs) are the most classic type of CPMs, which can be used in road safety evaluation stage and road operating and management stage. The most widely used CPM-GA in road safety evaluation is interactive highway safety design model (IHSDM) developed by AASHTO (2010). Although some researchers are more interested in the CPMs in road operating and management stage under real-time environment (Basso et al., 2018; Ba et al., 2017), the CPMs based on historical crash data and road geometric alignments are also very valuable to study, especially building CPM-GAs to predict annual crash number for target years.

The main purpose of this study is to examine a novel ensemble method for crash prediction by using road geometric alignments and traffic data to improve the predictive performance of tradition CPM-GAs and avoid the overfitting problem of machine learning algorithms. To be specific, the traditional CPM-GAs and machine learning CPM-GAs were combined to form more stable and more accurate CPM-GAs by using the new proposed ensemble rule. This study may contribute to current studies in the following four aspects. First, seven basic CPM-GAs (three traditional CPM-GAs and four machine learning CPM-GAs) were compared according to the model fitting and prediction accuracy, which had been rarely investigated in previous studies. Second, the sensitivity test of seven basic CPM-GAs was conducted, and the unstable characteristic of machine learning models was revealed in comparison with tradition models. Third, this study provided a new ensemble method of tradition CPM-GAs and machine learning CPM-GAs, which are believed to be beneficial for the final results of ensemble learning model. Fourth, the connecting "bridges" between tradition models and machine learning models were built in this study, which may shed light on more researches of CPMs in the ensemble learning framework.

2. Previous works

Various CPMs have been developed to date and can basically be divided into two categories: traditional methods and machine learning methods. The statistical models are the representation and the most popular approach of tradition CPMs, including Poisson, negative binomial (NB), Poisson-lognormal, zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB), generalized estimating equation (GEE), and various Bayesian models (Lord and Mannering, 2010). Recently artificial

intelligence technology is gaining momentum around the world, more and more researchers are getting interested in building CPMs by using support vector machine (SVM) (Ren and Zhou, 2011; Dong et al., 2015; Yu and Abdel-Aty, 2013), artificial neural network (ANN) (Zeng et al., 2016a; Zeng et al., 2016b), classification and regression tree (CART) and random forest (RF) (Wang et al., 2015). As a matter of fact, traditional CPMs can help researchers to be aware of the relationship between crash and crash-related factors but are only suitable for relative small size data not practical for big data or real-time data. Although machine learning CPMs have the problem of instability and overfitting, they always have higher prediction accuracy and efficiency than traditional CPMs. Therefore, the traditional CPMs and machine learning CPMs can be seem as two complementary methods and have their own strengths and weaknesses.

Among numerous types of CPMs, CPM-GAs are the most classic modeling method because traffic flow and road geometry have for years been recognized as contributing factors of crashes. Miaou (1994) studied the relationship between highway geometrics and accidents using NB models and found that NB regression models are more appropriate in instances where data are overdispersed. Milton and Mannering (1998) analyzed the relationship between annual accident frequencies and highway geometric and traffic characteristics by applying NB models. They concluded that the NB model is a powerful predictive tool. However, IHSDM has become the most popular CPM-GA after the publishing of *Highway Safety Manual* (AASHTO, 2010). Many studies made efforts to calibrate the crash prediction module of IHSDM and evaluate its transferability in different countries, such as the United States (Turner et al., 2012) and Canada (Persaud et al., 2012). Bauer and Harwood (2014) evaluated the safety effects of the combination of horizontal curvature and longitudinal grade on rural two-lane highways. CPMs for fatal-and-injury and property-damage-only crashes were built, and CMFs of five combinations of horizontal and vertical alignments were developed. Nevertheless, only a few studies utilized machine learning methods or other new methods to build CPM-GAs.

Ensemble learning (EL) method, which means the combination of multiple base models to form a stronger ensemble model, has been the state-of-the-art technology in machine learning field in recent years (Chen et al., 2017). Studies have shown that the EL can effectively improve models' prediction, generalizability, and robustness over a single model (Krawczyk et al., 2017). However, the concept was seldom applied or investigated in the traffic safety field. This study is expected to fill the gap by building ensemble learning CPM-GAs based on EL theory, and combing traditional CPM-GAs and machine learning CPM-GAs to form a higher accuracy model and stability model in the framework of EL.

2. Data preparation

2.1. Data

Data for this study were collected from 1,976 road segments with a total length of 371.4 km in length located on the following three freeways in China: Jingzhu Freeway, Kaiyang Freeway, and Yuegan Freeway. A total number of 3,970 crashes from 2009–2012 occurred on these freeways were collected. The data included crashes, traffic volumes (e.g., Annual Average Daily Traffic [AADT]), and geometric design characteristics.

The crash data were obtained from the Guangdong Provincial Freeway Administration (GDFA). Freeway geometric design data containing detailed road geometric characteristics were obtained from the Guangdong Provincial Communications Survey and Design Institute. First, the crash data were linked to the corresponding freeway segments with the help of kilometer markers. Next, road segments where there are toll station, tunnel, and bridges were eliminated. Then, three freeways were divided into homogeneous road segment, and each homogeneous road segment ends up with the minimum road unit according to their horizontal and vertical design geometric alignments (such as horizontal curves, transition curves, and tangents). To reduce the influence of black spots on the model prediction performance, the homogeneous road segments with anomalous crash frequency were removed by using the 3σ method (Francesca, Mariarosaria, & Gianluca, 2016). Finally, the data of basic road segments were extracted in this study. In addition, data are divided into three categories according to its location in different terrains (e.g., mountainous region, plain region, and mountainous hilly region). The descriptive statistics of variables are presented in Table 1.

2.2. Data normalization

Data normalization can improve the data fitting as well as prediction performance and is required for SVM and back propagation neural network (BPNN) models. The normalization was accomplished using the following equation in BPNN model (Li et al., 2008):

$$x_{n_i} = (x_i - \min(x_i))/(\max(x_i) - \min(x_i)) \tag{1}$$

Where, x_i is the variable i, $\min(x_i)$ is the minimum value of variable i, $\max(x_i)$ is the maximum value of variable i. The function "scale" in sklearn package of Python software (F. Pedregosa et al., 2011) was used before applying SVM models in this study for standardizing data to the data with the average value is 0 and the Standard Deviation is 1.

Table 1. Summary statistics of variables

Category	Variables	*M*	*SD*	Minimum	Maximum
Mountainous region	Response variable				
	Number of crashes in four years (crashes)	1.865	1.174	1	8
	Road characteristics				
	Length of segment (km)	0.174	0.108	0.051	0.845
	Traffic characteristics				
	AADT (pcu/day)	11761	2128	7465	14207
	Road geometric design				
	Horizontal curve radius (km)	1.417	1.530	0	0.800
	Absolute value of deflection angle (degree)	26.614	21.210	0	100.692
	Vertical curve radius (km)	12.107	16.162	0	200
	Slope gradient (%)	0.429	2.465	−5	5
	Slope length (km)	0.928	0.509	0.340	3.200
Plain region	Response variable				
	Number of crashes in four years	2.064	1.270	1	6
	Road characteristics				
	Length of segment (km)	0.263	0.173	0.053	1.098
	Traffic characteristics				
	AADT (pcu/day)	21942	1679	17601	25085
	Road geometric design				
	Horizontal curve radius (km)	2.626	2.747	0	8
	Absolute value of deflection angle (degree)	18.025	17.641	0	72.028
	Vertical curve radius (km)	19.271	23.960	0	150
	Slope gradient (%)	−0.001	0.920	−2.910	2.800
	Slope length (km)	0.843	0.299	0.400	2
Mountainous hilly region	Response variable				
	Number of crashes in four years	2.096	1.466	1	7
	Road characteristics				
	Length of segment (km)	0.232	0.137	0.056	1.197
	Traffic characteristics				
	AADT (pcu/day)	14124	2707	11078	19913
	Road geometric design				
	Horizontal curve radius (km)	2.280	2.746	0	9.800
	Absolute value of deflection angle (degree)	24.779	21.444	0	86.043
	Vertical curve radius (km)	15.833	27.387	0	300
	Slope gradient(%)	−0.092	2.019	−4	4
	Slope length (km)	0.653	0.324	0.146	1.728

Note. AADT = Annual Average Daily Traffic; pcu = passenger car unit.

2.3. Model fitting and prediction performance index

Mean absolute deviation (MAD) and mean squared predictor error (MSPE) proposed by Oh et al. (2003) were adopted to evaluate the model fitting and prediction performance. The measures of effectiveness (MOEs) are as follows (Oh et al., 2003):

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{2}$$

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \tag{3}$$

Where, n is the size of fitting or predicting sample, $\hat{y}_i$ is the estimated number of crashes of road segment i, and y_i is the observed number of crashes. The crash prediction model performance is better if the values of MAD and MSPE are smaller.

To evaluate the prediction accuracy and stability of ensemble learning model, two evaluation criteria were proposed in this study, including relative error change (REC) and relative variance change (RVC). The formulas of the two indexes are described as follows.

$$REC = (MAD_2 - MAD_1)/MAD_1 \tag{4}$$

$$RVC = (Variance_2 - Variance_1)/Variance_1 \tag{5}$$

Where, MAD_1 is the MAD of the traditional CPM, MAD_2 is the MAD of the ensemble learning CPM, REC is the relative error change, unit is %, $Variance_1$ is the variance of predicted results by machine learning CPM, $Variance_2$ is the variance of predicted results by ensemble learning CPM, and RVC is the relative variance change, unit is %. The ensemble learning model prediction accuracy is higher when REC is negative value, and the ensemble learning model stability is higher when RVC is negative value.

3. Methodology

3.1. Base model of ensemble learning: negative binomial regression

NB model assumes that the Poisson parameter follows a gamma distribution and NB regression model is usually given by the following (Miaou, 1994):

$$\text{Prob}(Y_i = y_i) = \frac{\Gamma(y_i + \phi)}{\Gamma(y_i + 1)\Gamma(\phi)} \left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i} \left(\frac{\phi}{\mu_i + \phi}\right)^{\phi} \tag{6}$$

$$\text{Expectation of } Y_i \text{ is } \mu_i = g(x_i) \tag{7}$$

$$\text{Variance of } Y_i \quad \text{is} \quad \text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\phi} \tag{8}$$

Where, y_i is the crash number of road segment i, Y_i is the dependent random variable following a NB distribution with the inverse dispersion parameter ϕ, x_i is the explanatory variables which related to crash at road segment i, and $g(x_i)$ is the link function for the model. In this study, $g(x_i)$ is described as follows:

$$\mu_i = (AADT_i \times L_i) \exp(\beta_0 + \beta_1 HC_i + \beta_2 HD_i + \beta_3 VC_i + \beta_4 SG_i + \beta_5 SL_i) \tag{9}$$

Where, $AADT_i$ is the annual average daily traffic for segment i, L_i is the length of segment in meter, HC_i is the horizontal curve radius for segment i, HD_i is the absolute value of deflection angle of horizontal curve for segment i, VC_i is the vertical curve radius for segment i, SG_i is the slope grade of segment i, SL_i is the slope length of segment i, and $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are regression coefficients.

3.2. Base model of ensemble learning: IHSDM model

IHSDM (IHSDM-US) was first developed in the United States and has been widely used to evaluate the safety performance of highway design alternatives. The CPMs in this software are described as follows:

$$N_{predicted} = N_{SPF_x} \times (CMF_{1x} \times CMF_{2x} \times \cdots \times CMF_{yx}) \times C_x \tag{10}$$

Where, $N_{predicted}$ is the predicted crash frequency on site type x(crashes/year), N_{SPF_x} is the predicted average crash frequency determined for base conditions with the SPF representing site type x (crashes/year), CMF_{yx} is the crash modification factors specific to site type x and variable y to study that can affect crash frequency, and C_x is the calibration factor to adjust the site type x for local conditions.

However, the real traffic environment and the road design standard are quite different between the United States and China, Hou (2014) developed the IHSDM (IHSDM-China) model for China and developed three SPFs for different terrain conditions (for more detail about the calculation of CMFs in IHSDM-China, please refer to Hou, 2014).

3.3. Base model of ensemble learning: Back propagation neural network

BPNN is the most popular and widely used algorithm for artificial neural network. Firstly, the output of hidden layer is calculated as (McClelland and Rumelhart, 1986):

$$H_j = f\left(\sum_{i=1}^{n} \omega_{ij} x_i - a_j\right) \quad j = 1, 2, \ldots, l \tag{11}$$

Where, H_j and f are the output of hidden layer and the incentive function of neurons, l is the neuron number of hidden layer, n is the neuron number of the input layer, ω_{ij} is the weight factor between input-layer and hidden layer, a_j is threshold value. Secondly, predicting value of the output layer is calculated as (McClelland and Rumelhart, 1986):

$$O_k = \sum_{j=1}^{l} H_j \omega_{jk} - b_k \quad k = 1, 2, \ldots, m \tag{12}$$

Where, b_k is threshold value, m is the neuron number of the output layer. Then according to the prediction error e_k calculated by the difference between predicted output and expected output, weight factor and threshold value are updated. Finally, the cycle of training calculation is judged by the termination conditions. In this study, the MATLAB software (The MathWorks, 2007) was used to build the BPNN. The specified parameter settings are as follows: input variable is road geometric alignments and

AADT, output is crash number; number of input is 4, number of output is 1, the number of hidden layer is 9, hidden layer function is sigmoid, output layer function is linear, maximum number of epoch is 1000, learning rate is 0.05, and accuracy is 0.0001 (the hidden layer units number is determined by the empirical equation, please refer to Hunter, Yu, Pukish, Kolbusz, & Wilamowski, 2012).

3.4. Base model of ensemble learning: support vector machine

SVM is a pattern recognition method based on statistical learning theory (Cortes and Vapnik, 1995). Assume the training input is defined as vectors $x(i) \in R^{In}$ for $i = 1, \ldots, N$, representing the important road geometric alignments variables and AADT. The training output is defined as $y(i) \in R^1$ for $i = 1, \ldots, N$, representing the crash frequency of roadway segments. The SVM maps $x(i)$ into a feature space $R^h (h > In)$ with higher dimension using $\phi(x(i))$ to linearize the nonlinear relationship between $x(i)$ and $y(i)$. Estimation function of $y(i)$ is

$$\hat{y} = f(x) = w^T \phi(x) + b \tag{13}$$

Where, $w \in R^h$ and $b \in R^1$ are coeffcients. The coefficients are derived by solving the following optimization problem (Cortes and Vapnik, 1995).

$$MinZ(w, \varepsilon, \xi, \xi^*) = \frac{1}{2} w^T w + C\left\{ \nu\varepsilon + \frac{1}{N} \sum_{i=1}^{N} \left(\xi_i + \xi_i^*\right) \right\} \tag{14}$$

subject to

$$w^T \emptyset(x(i)) + b - y(i) \leq \varepsilon + \xi_i \ \forall i = 1, \ldots, N \tag{15}$$

$$y(i) - w^T \emptyset(x(i)) - b \leq \varepsilon + \xi_i^* \ \forall i = 1, \ldots, N \tag{16}$$

$$\varepsilon \geq 0 \tag{17}$$

Where, ξ_i, ξ_i^* are slack variables; C is a regularization parameter, and ν is a second parameter, ε is the allowable error of each x(i). Slack variables ξ_i, ξ_i^* capture errors above ε and are penalized in the objective function through a regularization constant C. In this study, the SVM model with RBF kernel was built by Python (F. Pedregosa et al., 2011), and the grid-searching algorithm was used to find the best parameters. The settings are as follows: C is 1,10,100,1000 and gamma is 0.04, 0.2, 1, 5, 25.

3.5. Ensemble learning framework

EL, a state-of-the-art technology, has attracted growing attention in machine learning community (Polikar et al., 2012). EL combines the strengths of a pool of many simple base models to build a more robust

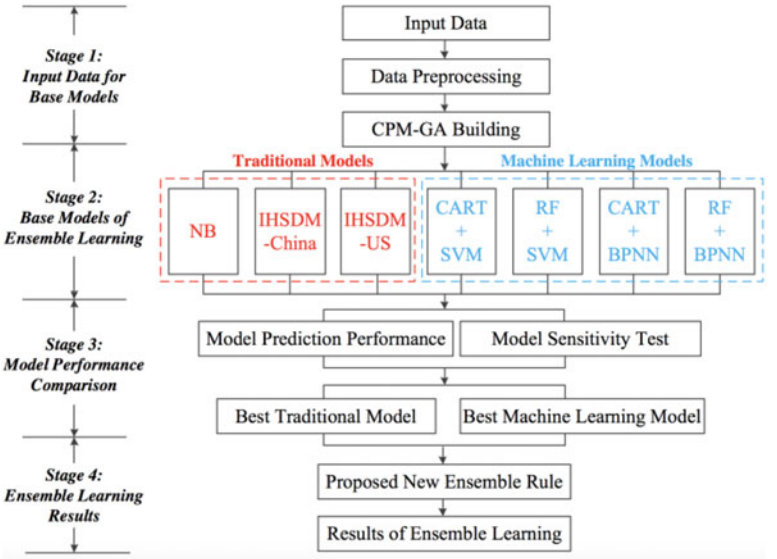

Figure 1. The framework of the proposed ensemble-learning for building crash prediction model.

combined model with higher prediction accuracy. According to "no free lunch" theorem (Wolpert, 2001), there is not a single classifier that is appropriate for all tasks because each algorithm has its own domain of competence. In this study, the framework of ensemble learning is shown in Figure 1.

The proposed ensemble learning approach to predict crash frequency in this study is different from existing ensemble learning methods:

- **Different data processing.** This study didn't use bagging or bootstrapping to randomly select data sets for base learners (Saha, Alluri, & Gan, 2015), because there is no distinct difference between these methods in this dataset.
- **Different base learners.** Traditional methods and machine learning methods were used as base learners in this framework, instead of only machine learning methods were adopted.
- **Different evaluating indexes.** Two evaluation aspects of crash prediction, model prediction accuracy and sensitivity test, were be taken into consideration for CPM-GAs for testing the transferability.
- **Different ensemble rules.** Different from the current ensemble learning rules, the new ensemble method was proposed to modify IHSDM model by using adjusting factor. The upper and lower bounds were set for controlling SVM's stability. The new ensemble rule can be expressed as the following formulas:

$$C_x = \frac{\overline{y_{SVM_x}}}{\overline{y_{IHSDM_x}}} \tag{18}$$

Where, C_x is the adjusting factor (new proposed concept, not the calibration factor of AASHTO) of IHSDM model in the terrain type x, $\overline{y_{SVM_x}}$ is the average predicted crash rate of IHSDM model in the terrain type x, $\overline{y_{IHSDM_x}}$ is the average predicted crash rate of SVM model in the terrain type x. This adjusting factor is used for capturing the other underlying and unobserved crash-related factors (*e.g.*, weather, dangerous behavior of drivers) and modifying the original predicted results to local conditions.

$$y'_{IHSDM_i} = y_{IHSDM_i} \times C_x, \quad i = 1, 2, \ldots, n \tag{19}$$

Where, i is the road segment, y'_{IHSDM_i} is the predicted crash rate of road segment i by modified IHSDM model, and y_{IHSDM_i} is the predicted crash rate of road segment i by original IHSDM model.

$$\sup(y_{SVM_i}) = y'_{IHSDM_i} + 0.25 \times y'_{IHSDM_i} \tag{20}$$

$$\inf(y_{SVM_i}) = y'_{IHSDM_i} - 0.25 \times y'_{IHSDM_i} \tag{21}$$

Where, $\sup(y_{SVM_i})$ is the upper bound of road segment i for SVM model, $\inf(y_{SVM_i})$ is the lower bound of road segment i for SVM model, and y_{SVM_i} is the predicted crash rate of road segment i for SVM model. Setting upper and lower bounds for SVM model can be useful to avoid abnormal predict crash value (e.g., very high crash rate) in original SVM model. Besides, 75%-125% was chosen for the change interval of modified results of IHSDM, which should be further discussed and verified in the future study.

$$y_{ensemble_i} = \begin{cases} \sup(y_{SVM_i}), y'_{SVM_i} > \sup(y_{SVM_i}) \\ y_{SVM}, \ \inf(y_{SVM_i}) \le y'_{SVM_i} \le \sup(y_{SVM_i}) \\ \inf(y_{SVM_i}), \ y'_{SVM_i} < \inf(y_{SVM_i}) \end{cases} \tag{22}$$

Where, $y_{ensemble_i}$ is the predicted crash rate of road segment i by ensemble learning method.

$$y_{IHSDM_i}, \ y_{SVM_i}, \ y'_{IHSDM_i}, \ y_{ensemble_i} \ge 0 \tag{23}$$

All predicted crash rate of road segment i by IHSDM model, SVM model, modified IHSDM model and ensemble learning model should be nonnegative number. The unit of crash rates is crashes per year per km.

4. Modeling results and discussion

4.1. Important variables selection results

CART and RF can be used as precursor to a more detailed regression model (Yu and Abdel-Aty, 2013). The importance of each variable is based

Table 2. Important variable selection results by Classification And Regression Tree (CART) and Random Forest (RF)

		Importance of variables	
Category	Selected important variables	CART	RF
Mountainous region	Slope gradient (%)	0.432[a]	0.388[a]
	Absolute value of deflection angle (degree)	0.265[a]	0.154[a]
	Slope length (m)	0.133[a]	0.243[a]
	Vertical curve radius (m)	0.099	0.120
	Horizontal curve radius (m)	0.072	0.094
Plain region	Vertical curve radius (m)	0.303[a]	0.290[a]
	Slope length (m)	0.265[a]	0.238[a]
	Slope gradient (%)	0.254[a]	0.274[a]
	Absolute value of deflection angle (degree)	0.128	0.125
	Horizontal curve radius (m)	0.049	0.073
Mountainous hilly region	Absolute value of deflection angle (degree)	0.393[a]	0.219[a]
	Slope length (m)	0.178[a]	0.259[a]
	Horizontal curve radius (m)	0.150[a]	0.151
	Vertical curve radius (m)	0.143	0.194[a]
	Slope gradient (%)	0.135	0.177

[a]the selected top three important variables by CART and RF, and AADT is the default important variables in each data set.

on Gini node purity, and a higher node purity value represents a higher variable importance (for more details about it, please refer to Breiman, 2001, and Wang et al., 2015). The response variable is crash number, and the predictor variables are five geometric alignments indexes (i.e., slope gradient, absolute value of deflection angle, slope length, vertical curve radius, and horizontal curve radius). It is worth mentioning that AADT is the important predictor variable, which is not considered in the important variable selection procedure. Due to the limitation of the original data of this study, the variables that are used for selecting important variables are not very ideal, and more road geometric alignment variables are suggested to be included in this procedure in the future studies. The settings in CART are as following: splitting criterion: Gini; maximum depth: 10, minimum leaf size: 10, minimum split size: 20. And the settings in RF are as following: bootstrap = true; criterion is mean square error, the number of tree is 100. Table 2 shows the importance of variables returned by CART and RF.

As is shown in Table 2, it is surprisingly found that the important variables chose by CART and RF are the same for freeways in mountainous region and plain region but are different for freeways in mountainous hilly region. This difference is presumably attributed to the different internal configurations of CART and RF algorithms (e.g., RF builds multiple decision trees and votes to make the final results). In fact, the slope gradient is the main factors that will greatly affect the crash rate in mountainous region because the maximum slope gradient even reaches 5%, and the gradient of 78 percentage of road segments is between –3% and 3%. Deflection angle is the major cause for mountainous region and

mountainous hilly region due to the complex and numerous horizontal curves in these areas. Besides, with the number of tangents and large-radius curves increasing in plain region, the vertical radius and slope lengths became crucial for crashes. Therefore, the comparison of model predictive performance need to be conducted to identify which variable selection method is better.

4.2. Comparison between prediction results of basic models

Based on the results of variable selection, four machine learning models were built: (1) CART + BPNN model, (2) RF + BPNN model, (3) CART + SVM model, and (4) RF + SVM model. The data set used in this study was randomly separated into two subsets, one is for training (60%, 70%, and 80% of samples), the other one is for testing (40%, 30%, and 20% of samples). Then, three traditional models were also built: (1) NB model, (2) IHSDM-China model, and (3) IHSDM-US model. Next, the seven base models of ensemble learning framework were compared with each other (the MOEs of all machine learning CPMs took the average value of training sizes of 60%, 70%, and 80%), as shown in Figure 2.

With regard to the two folds (MAD and MSPE criteria) of seven base model comparisons, four machine learning models have lower fitting and predictive errors for the training and testing data sets than the three traditional models. This demonstrates that machine learning algorithms for crash prediction may give a better approximation performance than NB models and IHSDM models, and SVM and BPNN models have good capacity for modeling nonlinear relationships between crash number and road geometric alignments (Zeng et al., 2016c).

Moreover, among three traditional CPMs, the NB model and IHSDM-China model outperforms than the IHSDM-US model. The reasonable explanation for this is that the SPFs should be developed for Chinese freeways and a calibration procedure should be conducted to adjust SPFs to reflect location conditions (AASHTO, 2010). SVM models are better performed than BPNN models in four machine learning models, indicating that the SVM models are particularly useful when the sample size is below 2,000 observations (Li et al., 2008). It is also noticeable that CART can be used as variable selecting procedure in developing crash prediction models (Yu and Abdel-Aty, 2013; Abdel-Aty and Haleem, 2011). After conducting sensitivity tests of RF + SVM, RF + BPNN, RF is not suggested for use in selecting important variables before building CPM-GA due to the instability of models, though RF were used as important variable selection in other studies (Abdel-Aty& Haleem, 2011; Haleem & Gan, 2013).

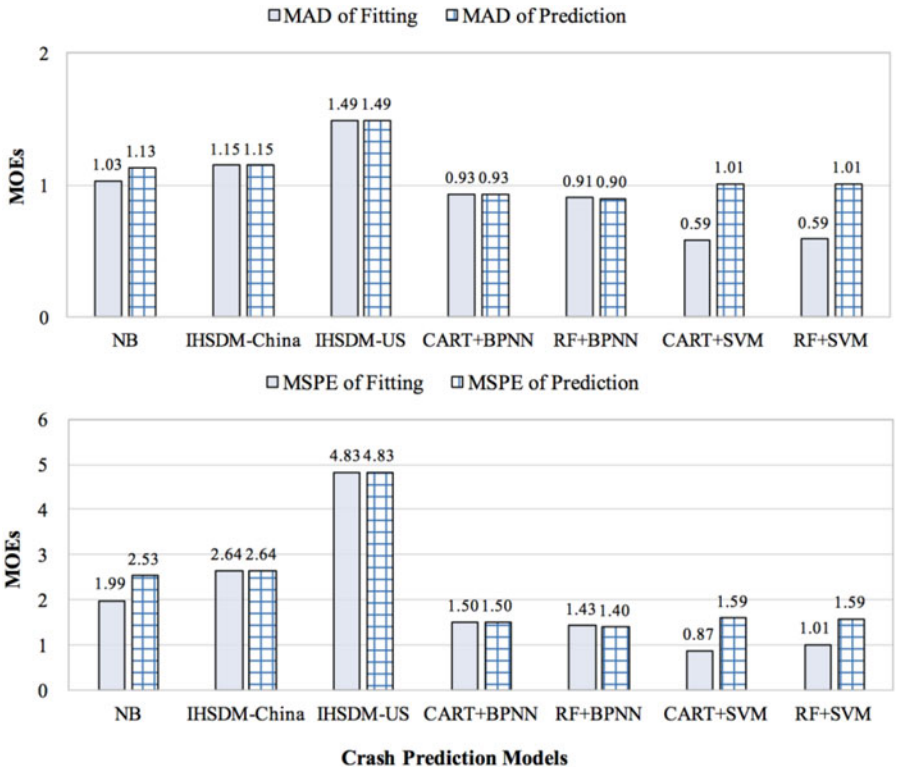

Figure 2. Comparisons of fitting and prediction performance between traditional Crash Prediction Model using Geometric Alignments and Traffic Data (CPM-GAs) and machine learning CPM-GAs.

4.3. Sensitivity analysis comparison of base models

To minimize the black-box problem of machine learning methods, the method proposed by Fish and Blodgett (2003) was adopted to analyze the sensitivity of machine learning models. Typical road segments which meet the basic conditions of IHSDM-China (for more details, please refer to Hou, 2014) were selected for sensitivity analysis (four road segments for each terrain conditions). After sensitivity analysis for seven basic CPMs were conducted, only IHSDM-China and CART + SVM were chosen due to other models' instability. And the sensitivity results of CART/RF + SVM and IHSDM-China (in mountainous region) are shown in Figure 3.

According to the results in Figure 3, the SVM crash prediction models are quite unstable when compared with IHSDM-China model in the sensitivity test, which are different from previous studies of SVM (Yu and Abdel-Aty, 2013; Li et al., 2008). This phenomenon indicates that when SVM models were applied in CPM-GA based on Chinese freeway crash data, the SVM models' results are tend to fluctuate within a certain range according to the varying predictor variables. This instability is contradictory

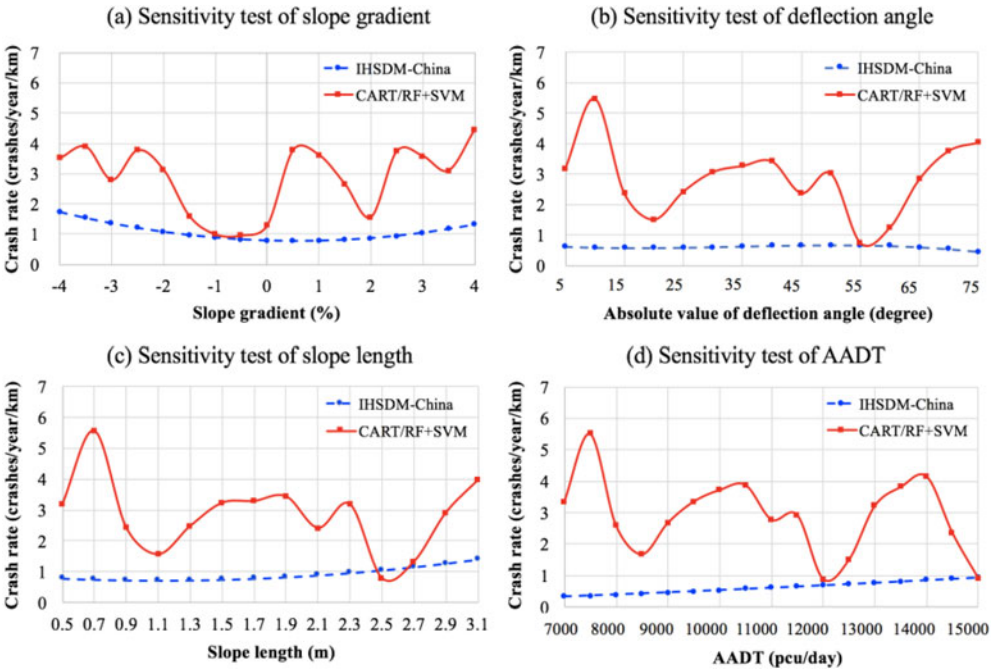

Figure 3. Comparison between Interactive Highway Safety Design Model in China (IHSDM-China) (blue dashed curve) model and Classification And Regression Tree + Support Vector Machine (CART+SVM) model (red curve) in the sensitivity analysis for different variables in mountainous region.

to the general relationships between the crash rate and road geometric alignments and traffic volume observed in previous studies (Miaou, 1994; Bauer and Harwood, 2014; Hou, 2014). The reason for SVM model generating this result maybe attributed that crash data have the stochastic feature in themselves and lack input data that are similar with the road segments of sensitivity test. However, when sensitivity analysis was conducted in IHSDM-China model, it shows regular and common relationships between crash rate and its related factors due to its internal functions forms.

It is also interesting to found that the fluctuate tendency of SVM model is almost the same for the sensitivity test of absolute value of deflection angle, slope length, and AADT, indicating that input variables have nearly the same effect on the output due to the radial basis function (RBF) in the SVM algorithm.

4.4. Ensemble learning results

As mentioned before, the new ensemble approach combining IHSDM-China model and CART + SVM model was applied and results of the ensemble learning CPM-GAs are positive in three data sets. The prediction accuracy and stability (model stability is measured by the variance of

Table 3. The prediction performance and sensitivity of ensemble learning CPM-GA

	IHSDM-China		CART + SVM		Ensemble learning			
Category	MAD	MSPE	MAD	MSPE	MAD	MSPE	REC (%)	RVC (%)
Mountainous region	3.22	63.74	2.71	51.73	2.87	55.93	−10.72	−35.47
Plain region	1.00	2.25	0.82	1.60	0.89	1.88	−11.15	−26.67
Mountainous hilly region	1.13	2.87	0.76	1.48	0.95	2.12	−15.91	−12.86

Note. CPM-GA = Crash Prediction Model using Geometric Alignments and Traffic Data; IHSDM = Interactive Highway Safety Design Model; CART = Classification And Regression Tree; SVM = Support Vector Machine; MAD = Mean Absolute Deviation; MSPE = Mean Absolute Deviation; REC = Relative Error Change; RVC = Relative Variance Change.

Relative error change of ensemble learning is compared with IHSDM-China model, and relative variance change is compared with CART + SVM model.

predict results) for basic models were both effectively improved, which is shown in Table 3.

From the results in Table 3, it is concluded that the proposed ensemble learning method in this paper can effectively increase prediction accuracy and stability at the same time. The average predict crash rate of the ensemble learning model is bigger than that of traditional IHSDM-China model and smaller than that of SVM model. The ensemble learning method can reduce the prediction error of IHSDM-China model by 10%–16% and reduce the variance of CART + SVM model by 12%-36%. This indicates that the ensemble learning model outperforms the tradition models and machine learning algorithms in aspects of model accuracy and stability, which confirms the major purpose of this study. Therefore, the proposed new ensemble learning CPM-GA can be a useful tool in road safety evaluation and road management. Given the fact that road geometry design greatly depends on the different terrains in China and a distinct difference of average crash rate in three trains (shown in Table 3), building different CPA-GAs based on three terrains (mountainous region, plain region, and mountainous hilly region) is recommended for predicting annual crash rate more accurately. Mountainous region usually has higher crash rate than the other terrains because of the presence of many long and steep longitudinal slopes (Meng et al., 2011).

5. Conclusions

Road safety evaluation is believed to have a promising future for improving the current traffic safety situation in Chinese freeway network. Therefore, stable and accurate CPMs are needed. CPM-GAs are the most classic modeling way in traffic safety analysis, which can be applied in road safety evaluation and road management. Given the fact that many previous studies mainly focused on the traditional statistical models to explore the relationship of crash rate and road geometric alignments and traffic volume, only a few studies revisit these relationships from the perspective of

machine learning algorithms and compare traditional CPM-GAs (NB and IHSDM) with machine learning CPM-GAs (BPNN and SVM) at the same time.

To fill this gap and inspired by the basic idea of EL, the ensemble learning framework and a new ensemble rule for connecting traditional CPM-GAs and machine learning CPM-GAs was proposed. The major objective of this study was to form an ensemble learning CPM with high stability and prediction accuracy. To accomplish this objective, the data for a total number of 3970 crashes occurred in three freeways in China were collected at first, and the input data were divided into three categories (mountainous region, plain region, and mountainous hilly region) to build CPM-GA, respectively. Next, the prediction accuracy and sensitivity of seven base CPMs (three traditional models and four machine learning models) were compared with each other at the same time. Finally, the best traditional model (IHSDM-China model) and the best machine learning method (CART + SVM model) were combined to form the final ensemble learning model, and the results show that:

- Whether CART or RF were employed as the procedure of critical variables selection, the fitting and prediction performance of SVMs and BPNNs was almost the same. However, CART is more recommended to be used before applying SVM models of CPM-GAs, because RF + SVM produced some negative values in the sensitivity tests in this study.
- Among seven base CPMs, machine learning models (SVM and BPNN) outperformed traditional models (NB and IHSDM) significantly in aspects of fitting and prediction accuracy than. However, machine learning CPMs suffered instability shortcomings in the sensitivity test; traditional methods had lower prediction accuracy than machine learning algorithms but have the advantages of model stability and transferability in different data sets.
- In comparison of three traditional CPM-GAs, NB model and IHSDM-China model has the highest prediction accuracy; CART + SVM has the highest prediction accuracy in contrast with other three machine learning models (CART + BPNN, RF + BPNN, RF + SVM). In sensitivity test of IHSDM-China and CART + SVM, the mean predicted value of crash rate in CART + SVM is relatively higher than that in IHSDM-China, and the results of CART + SVM showed the big variance and irregular characteristic with to the changing of predictor variables in some case.
- The new proposed ensemble learning CPM-GA was proved to be an effective way of reducing the variance of SVM models and passing the sensitivity test in this study. This ensemble method can improve the prediction accuracy of traditional IHSDM models as well. These results

may enlighten more innovation methodologies for building hybrid CPMs in the future.

Facing the challenges coming from artificial intelligence in traditional traffic safety analysis, the most important question is, with two major types of CPMs (traditional CPMs and machine learning CPMs) at present, which one is the best? Or could there be any connection between these two types CPMs? The answer given by this study is that, the two types CPMs are complementary with each other, and it's difficult to declare which type is better without considering its interest and data. There probably be a connection between them with the basic idea of ensemble learning. As mentioned before, traditional CPMs and machine learning CPMs have their own irreplaceable advantages, and ensemble of these two types CPMs may form a better comprehensive model with better predictive accuracy and transferability. The new ensemble learning CPM in this paper was examined to be useful to improve the accuracy and stability (avoiding the overfitting problems of SVM models) of basic CPM-GAs, which realized the final purpose of this study.

However, the ensemble learning CPM still has the issue of black-box, and future investigation can focus on verifying this new ensemble approach in different datasets. The authors hope that the idea presented in this paper may shed light on more researches for hybrid CPMs which connect traditional and machine learning algorithms in traffic safety field in the future.

Acknowledgement

This study was supported by the National Natural Science Fund (No. 71701055) in China. The authors want to specially thank Professor Zheng and Dr. Hou who gave many precious comments on this paper.

References

AASHTO. (2010). *Highway safety manual. American association of state highway transportation officials* (pp. 1500).

Abdel-Aty, M., & Haleem, K. (2011). Analyzing angle crashes at unsignalized intersections using machine learning techniques. *Accident Analysis and Prevention*, *43*(1), 461. doi: 10.1016/j.aap.2010.10.002

Basso, F., Basso, L. J., Bravo, F., & Pezoa, R. (2018). Real-time crash prediction in an urban expressway using disaggregated data. *Transportation Research Part C Emerging Technologies*, *86*, 202–219. doi:10.1016/j.trc.2017.11.014

Ba, Y., Zhang, W., Wang, Q., Zhou, R., & Ren, C. (2017). Crash prediction with behavioral and physiological features for advanced vehicle collision avoidance system. *Transportation Research Part C: Emerging Technologies*, *74*, 22–33. doi:10.1016/j.trc.2016.11.009

Bauer, K. M., & Harwood, D. W. (2014). Safety effects of horizontal curve and grade combinations on rural two-lane highways. *Transportation Research Record: Journal of the Transportation Research Board*, *14*, 37–49. doi:10.3141/2398-05

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chen, X., Zahiri, M., & Zhang, S. (2017). Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach. *Transportation Research Part C Emerging Technologies*, *76*, 51–70. doi:10.1016/j.trc.2016.12.018

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

Dong, N., Huang, H., & Zheng, L. (2015). Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects. *Accident Analysis and Prevention*, *82*, 192–198. doi:10.1016/j.aap.2015.05.018

Francesca, R., Mariarosaria, B., & Gianluca, D. (2016). Safety performance functions for crash severity on undivided rural roads. *Accident Analysis and Prevention*, *93*, 75–91.

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*(3), 244–254. doi:10.1057/palgrave.jt.5740081

Hou, Q. Z. (2014). Traffic Accident Prediction Model Based on IHSDM framework. Doctoral dissertation, Harbin Institute of Technology. (in Chinese with English abstract).

Haleem, K., & Gan, A. (2013). Effect of driver's age and side of impact on crash severity along urban freeways: A mixed logit approach. *Journal of Safety Research*, *46*, 67–76. doi:10.1016/j.jsr.2013.04.002

Hunter, D., Yu, H., Pukish, M. S., III, Kolbusz, J., & Wilamowski, B. M. (2012). Selection of proper neural network sizes and architectures: a comparative study. *IEEE Transactions on Industrial Informatics*, *8*(2), 228–240. doi:10.1109/TII.2012.2187914

Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion*, *37*(C), 132–156. doi:10.1016/j.inffus.2017.02.004

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), 291–305. doi:10.1016/j.tra.2010.02.001

Li, X., Lord, D., Zhang, Y., & Xie, Y. (2008). Predicting motor vehicle crashes using support vector machine models. *Accident Analysis and Prevention*, *40*(4), 1611. doi:10.1016/j.aap.2008.04.010

Miaou, S. P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Analysis Analysis and Prevention*, *26*(4), 471–482. doi:10.1016/0001-4575(94)90038-8

Milton, J., & Mannering, F. (1998). The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. *Transportation*, *25*(4), 395–413.

Miaou, S. P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Analysis and *Prevention*, *26* (4), 471–482. doi:10.1016/0001-4575(94)90038-8

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing. Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Oh, J., Lyon, C., Washington, S., Persaud, B., & Bared, J. (2003). Validation of FHWA crash models for rural intersections: Lesions learned. *Transportation Research Record: Journal of the Transportation Research Board*, *1840*(1), 41–49. doi:10.3141/1840-05

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: machine learning in python. *J. Machine Learning Research*, *12*, 2825–2830.

Persaud, B., Saleem, T., Faisal, S., Lyon, C., Chen, Y., & Sabbaghi, A. (2012). Adoption of highway safety manual predictive methodologies for Canadian highways. In: Paper Presented at the Proceedings from the 2012 Conference of the Transportation Association of Canada, Fredericton, New Brunswick.

Polikar, R., Zhang, C., & Ma, Y. (2012). Ensemble machine learning: methods and applications.

Ren, G., & Zhou, Z. (2011). Traffic safety forecasting method by particle swarm optimization and support vector machine. *Expert Systems with Applications*, *38*(8), 10420–10424.

Saha, D., Alluri, P., & Gan, A. (2015). Prioritizing highway safety manual's crash prediction variables using boosted regression trees. *Accident Analysis & Prevention*, *79*, 133–144.

Traffic Management Bureau of Ministry of Public Security. (2013)., *National Highway Network Plan (2013–2030)*. China: Traffic Management Bureau of Ministry of Public Security.

Traffic Management Bureau of Ministry of Public Security. (2014)., *Road traffic accidents statistics of PRC Beijing*, China: Traffic Management Bureau of Ministry of Public Security.

Turner, D. S., Jones, S., Jr., Lou, Y., Brown, D. B., & Smith, R. K. (2012). Implementation of the AASHTO Highway Safety Manual. UTCA Report (10404).

The MathWorks, Inc. (2007). *MATLAB Programming*. Natick, MA: The MathWorks, Inc.

Wang, J., Zheng, Y., Li, X., Yu, C., Kodaka, K., & Li, K. (2015). Driving risk assessment using near-crash database through data mining of tree-based model. *Accident Analysis & Prevention*, *84*, 54–64. doi:10.1016/j.aap.2015.07.007

Wolpert, D. H. (2001). The supervised learning no-free-lunch theorems, in: Proceedings of the 6th online world conference on soft computing in industrial applications, 2001, pp.25–42.

Meng, X. H., Guan, Z. Q., & Zheng, L. (2011). Safety evaluation of mountainous expressway based on geometric alignment indexes. *China Journal of Highway and Transport*, *24*(2), 103–108. (In Chinese with English abstract).

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis and Prevention*, *51*, 252–259. doi:10.1016/j.aap.2012.11.027

Zeng, Q., Huang, H., Pei, X., Wong, S. C., & Gao, M. (2016a). Rule extraction from an optimized neural network for traffic crash frequency modeling. *Accident Analysis and Prevention*, *97*, 87. doi:10.1016/j.aap.2016.08.017

Zeng, Q., Huang, H., Pei, X., & Wong, S. C. (2016b). Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. *Analytic Methods in Accident Research*, *10*, 12–25. doi:10.1016/j.amar.2016.03.002

Zeng, Q., Huang, H., Pei, X., & Wong, S. C. (2016c). Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. *Analytic Methods in Accident Research*, *10*, 12–25. doi:10.1016/j.amar.2016.03.002