Accident Analysis and Prevention 135 (2020) 105343

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Comparison of different models for evaluating vehicle collision risks at upstream diverging area of toll plaza

Lu Xing[a], Jie He[a,*], Ye Li[b], Yina Wu[c], Jinghui Yuan[c], Xin Gu[a]

[a] School of Transportation, Southeast University, 2 Si pai lou, Nanjing, 210096, PR China
[b] School of Traffic and Transportation Engineering, Central South University, Changsha, 410075, PR China
[c] Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, Florida 32816-2450, USA

ARTICLE INFO

Keywords:
Collision risk
Trajectory data
Toll plaza
Diverging area
Logistic regression
Non-Parametric model

ABSTRACT

Toll plazas with both Electronic Toll Collection (ETC) lane(s) and Manual Toll Collection (MTC) lane(s) could increase crash risks especially at upstream diverging areas because of frequency lane-change behaviors. This study develops the logistic regression (LR) model and five typical non-parametric models including, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees (DT), and Random Forest (RF) to examine the relationship between influencing factors and vehicle collision risk. Based on the vehicle trajectory data extracted from unmanned aerial vehicle (UAV) videos using an automated video analysis system, the unconstrained vehicle motion's collision risk can be evaluated by the extended time to collision (ETTC). Results of model performance comparison indicate that not all non-parametric models have a better prediction performance than the LR model. Specifically, the KNN, SVM, DT and RF models have better model performance than LR model in model training, while the ANN model has the worst model performance. In model prediction, the accuracy of LR model is higher than that of other five non-parametric models under various ETTC thresholds conditions. The LR model implies a pretty good performance and its results also indicate that vehicle yields the higher collision risk when it drives on the left side of toll plaza diverging area and more dangerous situations could be found for an ETC vehicle. Moreover, the vehicle collision risks are positively associated with the speed of the following vehicle and the angle between the leading vehicle speed vector and X axis. Furthermore, the results of DT model show that three factors play important roles in classifying vehicle collision risk and the effects of them on collision risk are consistent with the results of LR model. These findings provide valuable information for accurate assessment of collision risk, which is a key step toward improving safety performance of the toll plaza diverging area.

1. Introduction

Toll roads have been utilized to improve traffic efficiency via economic approaches to controlling demand, while toll plaza is a physical structure for collecting charges with several tollbooths on roads. However, despite of the benefits of constructing toll plazas, many previous studies reported that tollbooths may increase crash risks (Abdelwahab and Abdel-Aty, 2002; Carroll, 2016; Mckinnon, 2013; Saad et al., 2018), especially at the upstream areas of toll plazas. For example, Abuzwidah (2011) found that the diverging areas had 82% higher risk of traffic crashes than merging areas after toll, since upstream areas of toll plazas have limited space, complicated lane configurations and different toll collection types, which may increase drivers' confusions. Meanwhile, drivers need to diverge into the target toll collection lanes before the end of the diverging area, which could lead to more aggressive deceleration, acceleration, or lane-changing behaviors (Saad et al., 2018).

It is worth noting that the market penetrate rate of electronic payments may be less than 100%, thus both cash payments and electronic payments should be utilized for toll plaza. Drivers using the electronic payments could drive without stop on electronic toll collection (ETC) lanes, while drivers using cash need to decelerate to complete stops on manual toll collection (MTC) lanes. The toll plazas that have both MTC lanes and ETC lanes can be classified into two types based on different layouts of ETC lanes and MTC lanes (Abuzwidah and Abdel-Aty, 2015; Xing et al., 2019). As shown in Fig.1, the first type is "traditional mainline toll plaza (TMTP)", which has both ETC lanes and MTC lanes. The other type is "hybrid mainline toll plaza (HMTP)", in which

* Corresponding author.
E-mail addresses: luxing@seu.edu.cn (L. Xing), hejie@seu.edu.cn (J. He), yelicsu@csu.edu.cn (Y. Li), jessicawyn@Knights.ucf.edu (Y. Wu), jinghuiyuan@Knights.ucf.edu (J. Yuan), guxin0307@126.com (X. Gu).

https://doi.org/10.1016/j.aap.2019.105343
Received 17 May 2019; Received in revised form 16 September 2019; Accepted 17 October 2019
Available online 22 November 2019
0001-4575/

Fig. 1. Different designs of toll plaza area. (a) traditional mainline toll plaza; (b) hybrid mainline toll plaza. (Source: Xing et al. (2019), Central Florida Expressway Authority& Google map)

vehicles using electronic payment method are completely separated from the vehicles using cash at upstream area of toll plaza. The first type is prevalent in China, Korea, and England, while the latter type is widely utilized in the United States.

For these two types of toll plazas, the TMTP is more dangerous due to more complicated toll plaza design as well as traffic conditions (Abuzwidah and Abdel-Aty, 2015; Xing et al., 2019). In the diverging area of TMTP, the MTC and ETC vehicles are mixed together and have to change lanes to head for their target toll collection lanes in the same area, which results in the more frequently crossing behaviors and the speed variations (Abdelwahab and Abdel-Aty, 2002). Additionally, end parts of diverging areas do not have lane marking, which also causes more unordered and unpredictable diverging behaviors (Carroll, 2016; Mckinnon, 2013; Xing et al., 2019). Moreover, ETC vehicles need to slow down to 20 km/h for passing the toll collection lane, which could increase the rear-end collision risk. Therefore, it is worthy to investigate vehicle's safety in this area.

Different types of data have been utilized for traffic safety analysis. In general, crash data could be analyzed together with aggregated traffic data from infrastructure to establish the relationship between accidents and contributing factors (de Oña et al., 2013; Huang et al., 2018; Yuan and Abdel-Aty, 2018), which includes loop detection, Microwave Vehicle Detection System (MVDS), etc. In recent years, with the development of technologies, there is a clear trend of using trajectory data to conduct traffic analysis, which could be obtained from floating vehicles or videos. The trajectory data provide more detailed microscope information related to drivers' behavior for each vehicle that could be employed for traffic safety analysis (Guo et al., 2016; Laureshyn et al., 2010; Mahmud et al., 2017; Sayed and Zein, 1999; Weng et al., 2014; Wu et al., 2019, 2019b; Zheng et al., 2014). Regarding models in safety analysis, considerable crash risk models have been proposed in previous studies to reveal influence factors and predict accident probability. Most of these studies utilized the parametric technique such as logistic regression (LR) model to establish relationship between crash risks (dependent variable) and various factors (independent variables) (Abdel-Aty et al., 2012; Weng and Meng, 2011; Wu et al., 2018; Xu et al., 2015; Yuan et al., 2018; Meng and Weng, 2011). The outcomes of LR model can predict the probability of crash and estimate the marginal effect of each explanatory variable. It has good theoretical interpretability and clear calculation construction. However, the limitation of LR is apparent that the assumptions for data distribution have to be satisfied, otherwise the incorrect inferences could be produced. Besides, the pre-defined underlying relationships between dependent and independent variables are constrained by the model itself, which is another important limitation (Chang and Chien, 2013; Delen et al., 2006; Mussone et al., 1999; Xu et al., 2013). Non-parametric models including K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and so on. (Abdelwahab and Abdel-Aty, 2007; Abellán et al., 2013; Das et al., 2009; Li et al., 2012; Siddiqui et al., 2012) are also widely proposed for predicting vehicle collisions. Weng and Meng (2012) developed a DT using the classification and regression tree (CART) algorithm to analyze the relationship between the risky driving behavior and its influencing factors at work zones. De Oña et al. (2013) and Huang et al. (2018) also used DT to investigate the relationship between crash severity and various risk factors on rural highways and mountainous freeways respectively. Dong et al. (2015) developed SVM models in traffic analysis zones level crash risk analysis with consideration of spatial correlations. The models have been adopted to predict the various merging behaviors at expressway on-ramp bottlenecks (Wang et al., 2017). The RF model has been also widely employed to predict crash and investigate the significant impacts of various factors on vehicle crash likelihood (Shi and Abdel-Aty, 2015; Siddiqui et al., 2012). Compared to LR model, non-parametric model usually can provide a high level of prediction accuracy, also they have no data distribution assumptions and the functional form between independent and explanatory variables is well defined (Xu et al., 2015).

Meanwhile, the parametric technique and non-parametric technique were compared in many previous studies. Some studies have suggested that the non-parametric models are suitable for safety analysis, because they could provide high prediction accuracy, avoid the inherent problems occurred in the LR model, and display the relationship between crash and various factors using only a few essential variables with a brief graphic (Chang and Chien, 2013; Huang et al., 2018; Jung et al., 2016; Weng and Meng, 2012). However, other studies argued that the results of non-parametric models could have a weak interpretation of results, also is less useful in examining the marginal effects of factors, which can provide valuable information for traffic engineers to establish the priorities for risk mitigation. Hence, the LR model is more practical and convenient for safety analysis (Weng and Meng, 2011; Xu et al., 2015). Moreover, some studies compared these two techniques about model performances (Kuhnert et al., 2000; Pakgohar et al., 2011). For example, Ali et al. (2019) applied these two techniques to detect near-crashes on freeways, the results suggested that LR model provided a good fit of the input data and can detect near-crashes with outstanding discrimination ability. However, non-parametric models

had better performance when considering time factor, DT and ANN had higher accuracy on near-crashes detection compared to the KNN.

In conclusion, these two techniques could perform differently with different research objectives, datasets, scenarios, etc. Therefore, the present study intends to compare the performances of various non-parametric models and LR model for micro-level safety analysis, i.e., evaluating vehicle safety at the upstream toll plaza diverging area using microscopic trajectory data. To achieve this objective, the trajectory data are collected via unmanned aerial vehicle (UAV) and extracted by an automated video analysis system, and vehicles' collision risk is computed by extended time to collision (ETTC), which can be adopted for evaluating unconstrained vehicle motion's collision risk. Three different values of ETTC threshold are set for model validation. Up to five different non-parametric models, including KNN, SVM, ANN, DT, and RF, are employed as well as the LR model for comparison. This study contributes to traffic safety analysis at the upstream diverging area of toll plaza along the following directions: (i) evaluate the collision risk of unconstrained vehicle motions at toll plaza diverging area; (ii) employ parametric and non-parametric models based on microscopic vehicle trajectory data and the best modeling approach for the traffic safety analysis at toll plaza diverging area is suggested; (iii) evaluate the safety performance of vehicles and examine the impacts of the different factors on vehicle collision risk at toll plaza diverging area.

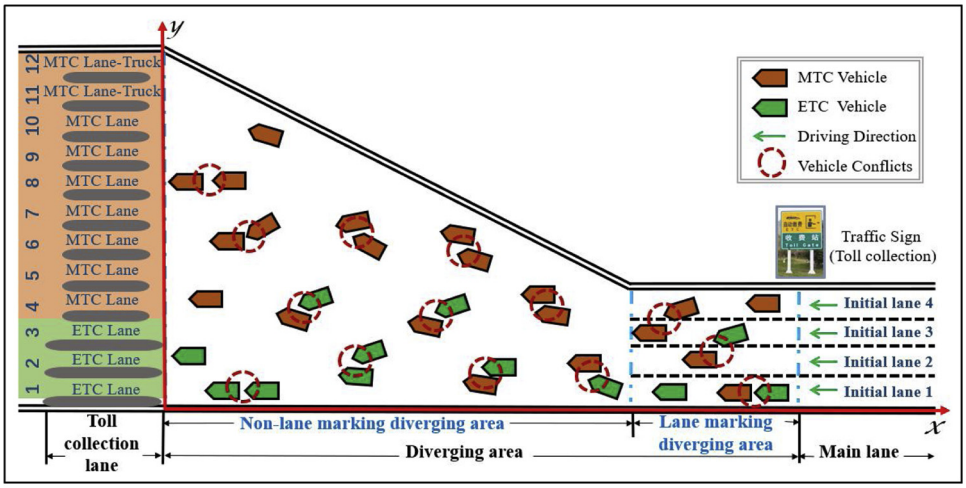
2. Data source

To investigate vehicle collision risks at the upstream area of toll plaza, data were collected at a toll plaza area on G42 freeway in Nanjing, China. G42 is an east-west direction toll way with eight lanes for each direction, serving as a major corridor in the northeast area of Nanjing city. Fig. 2 displays the layout of the diverging area in this study site. The toll plaza diverging area was defined as consisting of two parts: lane marked diverging area and non-lane marked diverging area (Xing et al., 2019). As shown in Fig. 2, there are four lanes of lane marked diverging area (60 m), and a long diverging area without any lane marks (300 m). In the toll collection area that is adjacent to the end of diverging area, there are 3 ETC lanes located on the left side, while 9 MTC lanes located on the right side. Vehicles from main lanes need to diverge into one of the twelve toll collection lanes. Moreover, vehicles need to drive to the lanes with the corresponding payment methods. For example, the vehicle installed ETC tag need to drive to the leftmost 3 ETC lanes and drivers using cash cannot use these 3 toll collection lanes. The phenomenon may lead to higher crash risks for toll plazas.

The UAV collected the traffic video with mounted cameras during the peak period on March 17th, 2018 and it was a sunny, windless day. It provides a stable video with the quality of 4 K ultra high definition and 30 frames per second (fps). About 1.5 h videos were recorded and 50 min videos were selected for analysis. The equivalent hourly traffic flow rate of this diverging area ranged from 1050 vph to 1740 vph (calculated as 6 times of the 10-minute volume). With the collected video, a procedure of video-data processing was conducted, including stabilization, features detection and tracking, coordinate transformation, and error elimination. The automated video analysis system has the higher detecting and tracking accuracy compared to the conventional video analysis system, the data collection and processing are detailed in our previous study, (Xing et al., 2019; Wu et al., Under Review). Based on the data processing, 1031 cars were tracked, and their trajectories were obtained. Further, the vehicles are deleted having no vehicle surround it in the total diverging area, and 1016 vehicles' trajectories are applied for safety analysis finally.

The trajectory data contains various dynamic information, such as frame ID, time ID, position, etc. The vehicles' toll collection type (MTC& ETC) were extracted manually. Based on the trajectory data, other information can be obtained, including the vehicle speed, the angle between vehicle speed vector and the X-axis (Theta angle), the vehicle's initial lane, the vehicle's target toll collection lane, the distance

(a)



(b)

Fig. 2. Layout of diverging area in the study site (Xing et al., 2019) (a) image collected by UAV; (b) illustration of layout and potential conflicts.

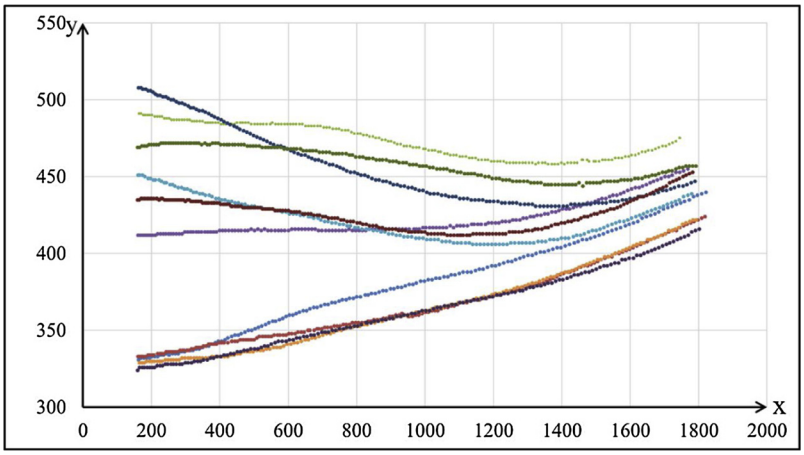


Fig. 3. The sample of 10 vehicles' trajectory data. (Note: pixel coordinates on the video image).

between two vehicles' centroids, the traffic flow volume, etc. As shown in Fig. 2, there are four lanes of main lane and all vehicles in the diverging area come from one these lanes, so they are called as vehicle's initial lane in this study. Moreover, there are also twelve toll collections and the sequence numbers ascend from the left side to the right side. The abundant trajectory data provides useful details for safety analysis. A sample of 10 vehicles' trajectories are displayed in Fig. 3. It is apparent that all spatially and temporally dynamic details could be included, and ETC vehicles and MTC vehicles have remarkably different trajectories.

3. Methodology

3.1. Modeling approach

3.1.1. Logistic Regression (LR) model

The LR model has been widely utilized for investigating the relationship between a binary response and independent variables (Abdel-Aty et al., 2012; Weng and Meng, 2011; Wu et al., 2018; Xu et al., 2015; Yuan et al., 2018). The conditional probability of vehicle potential collision ($p(y_n)$) can be calculated by Eq. (1):

$$p(y_n) = \frac{e^{g(y_n)}}{1 + e^{g(y_n)}} \tag{1}$$

The LR model can be expressed in Eq. (2):

$$g(y_n) = ln\frac{p(y_n)}{1 - p(y_n)} = \boldsymbol{\beta x} + \varepsilon_n = \beta_{0n} + \beta_{1n}X_{1n} + \beta_{2n}X_{2n} + \ldots + \beta_{kn}X_{kn} + \varepsilon_n \tag{2}$$

Where n indicates the total number of observations; $\boldsymbol{x}$ is a set of the independent variables and $\boldsymbol{\beta}$ is a set of corresponding parameters. The term k is the number of independent variables, and ε_n is the error term with a normal distribution with mean zero (Wu et al., 2018; Xing et al., 2019).

3.1.2. Non-parametric models

Five different non-parametric models are employed for safety analysis. The data of influencing factors are input into these models and the output is the prediction accuracy of collision risks.

3.1.2.1. K-Nearest neighbor (KNN) model. The KNN is a typical machine learning algorithm which takes inputs of the K closest training examples in the feature space and outputs a class membership or property value. For classification problem, we find the class most of the closest K examples belong to and use it as the output class. To implement the KNN algorithm, the value of K and a distance function need to be determined. In this study, K is a major parameter of KNN, the best value of it will be ascertained via accuracy curves. The Euclidean distance is utilized as the distance function. It is a commonly used distance metric of KNN, which can be described as a physical distance between two points (Altman and Altman, 2012; Iranitalab and Khattak, 2017), which can be defined as:

$$d_{ij} = \left(\sum_{K=1} (x_{iK} - x_{jK})^2\right)^{1/2} \tag{3}$$

where d_{ij} is the distance between observations i and j; x_{iK} and x_{jK} and are values of the Kth variable for observations i and j.

3.1.2.2. support vector machines (SVM) model. The SVM is a supervised learning model for classification and regression analysis. The basic idea of SVM is to find a best separation of a hyperplane having the largest distance to the closest data point of any class. The best separation will have the largest margin and smallest classification errors (Cortes and Vapnik, 1995; Li et al., 2012). In the present study, the C-classification SVM is employed, with the parameter gamma for the kernel function needing to be determined. Readers could refer to the study of Kecman (2005)for more details about C-SVM and the affiliated kernel function and its parameters.

3.1.2.3. artificial neural networks (ANN) model. The original goal of ANN is to mimic a human brain for solving problems. It has become a powerful model for classification and regression analysis. A variety of ANN has been proposed, including back-propagation neural networks, convolutional neural networks, long short-term memory neural networks and so on. In this research, the back-propagation neural networks are utilized for safety evaluation. The hidden layer size of the ANN is a key parameter, which will be determined by accuracy curves (Abdelwahab and Abdel-Aty, 2002, 2007; Porto-pazos et al., 2011).

3.1.2.4. decision trees (DT) model. The DT is a common-used supervised learning method for classification and regression. It utilizes a tree-like model to make decisions based on simple rules inferred from the data features. The DT is a simple model for understanding and visualization, which is an advantage among all the machine learning models. The maximum depth of a tree is the parameter needing to be determined. The DT is categorized as the classification tree and the regression tree, corresponding to the categorical and continuous target variables respectively (Breiman et al., 1984).

In this study, the CART (Classification and Regression Trees) is used to build DTs, which is developed by Breiman et al. (1984) and has been most commonly used in safety analysis. The CART uses a decision tree to solve classification and regression problems, and relationships between factors and vehicle collision risk can be identified by splitting a

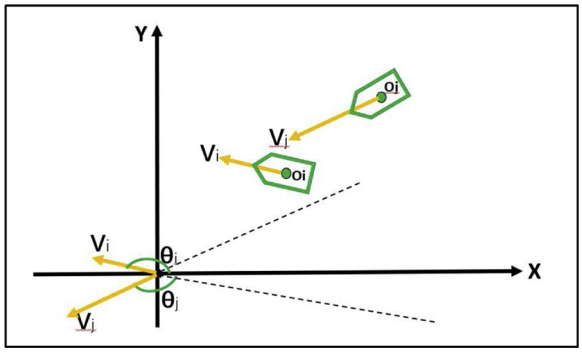


Fig. 4. Illustration of two approaching vehicles' positions in the coordinate system.

large dataset into more homogenous subsets during the classification process using the Gini index (Huang et al., 2018).

The Gini index is defined as:

$$\text{gini(C)} = 1 - \sum_j p^2(C = c_j) \tag{4}$$

$$\text{GIx(C, X)} = \text{gini(C|X)} - \text{gini(X)} \tag{5}$$

where $\text{gini(C|X)} = \sum_j p(x_t)\text{gini(C|}\ \text{X}= x_t)$ and X is another variable. The best split can be obtained by the minimum GIx(C, X).

3.1.2.5. random forest (RF) model. The RF is an ensemble learning method which consists of a collection of tree-structured classifiers and each tree casts a unit vote for the most popular class at input. It constructs a multitude of DTs and picks the best split among a random subset of features. It is different from the DT that chooses the best split among all features. The randomness always yields the better performance of RF models. The number of trees in the forest, defined as the number of estimators, is the key parameter for determination. For more details about these non-parametric models, readers can refer to the reference of machine learning (Breiman, 2001; Cheng et al., 2019).

3.2. Extended time-to-Collision (ETTC)

In order to evaluate vehicle collision risk, the surrogate safety measure (SSM) is applied to establish the relationship between conflict probability and trajectory data (Perkins and Harris, 1968; Sayed et al., 2013; Zheng et al., 2014). Traditional SSM indicators, such as TTC and deceleration rate avoid crash (DRAC), usually cannot be utilized directly, because these indicators are mainly for rear-end conflict analysis. Most studies calculated the conventional TTC based on the assumption that the consecutive vehicles are in the same traffic lane or their trajectories cross at a right angle (Behbahani and Nadimi, 2015; Laureshyn et al., 2010; Mahmud et al., 2017). In our study, vehicles in the diverging area will approach each other at any angle, especially in the non-lane marked diverging area. Thus, an appropriate indicator should be screen out, which can be not only applied to all kinds of conflicts analysis, but also easily extracted out anywhere in a large area. In this study, an extended TTC (ETTC) is utilized to estimate the individual vehicle collision risk. It is an extension of the conventional TTC to the more general case of two-dimensional movement, and more widely used for various traffic situations (Ward et al., 2015; Xing et al., 2019).

The TTC is defined as the time required for two vehicles in the same lane to collide if they do not change their states (change speeds or lanes) (Hayward, 1972). The TTC of a following vehicle j at time step t with respect to the leading vehicle i can be expressed as follows:

$$TTC_j(t) = \begin{cases} \dfrac{X_i(t) - X_j(t) - L_i}{V_j(t) - V_i(t)}, & \text{if } V_j(t) > V_i(t) \\ \infty, & \text{if } V_j(t) \le V_i(t) \end{cases} \tag{6}$$

where X and V denote the positions of vehicle's the most front point and vehicle speeds, respectively, and L denotes the length of vehicle. It assumes that the X of leading vehicle is larger than that of following vehicle. According to the definition, the larger value of TTC could provide more time for drivers to change their states and avoid a crash, while the smaller value may cause a higher probability of risks (Li et al., 2014, 2016, 2017a, 2017b).

The ETTC was proposed by Ward et al. (2015) and Xing et al.(2019) developed it using vehicle trajectory data, aiming to evaluate conflict risk of unconstrained vehicle motion. The ETTC can be expressed as follows:

$$ETTC_j = -\frac{D_{ij} - 0.5L_i - 0.5L_j}{\frac{1}{d_{ij}}(\boldsymbol{O}_i - \boldsymbol{O}_j)^{\mathrm{T}}(\boldsymbol{V}_i - \boldsymbol{V}_j)} \tag{7}$$

Where D_{ij} is the distance between two vehicles' centroids, d_{ij} is the distance between two closest points of two vehicles, L_i and L_j are lengths of the leading and following vehicles, respectively. As shown in Fig. 4, vehicle i is the leading vehicle and vehicle j is the following vehicle. The $\boldsymbol{O}$ and $\boldsymbol{V}$ are two-dimensional coordinates and speed vectors of vehicle's centroid. The details of ETTC's derivation can be referred in studies of Ward et al. (2015) and Xing et al.(2019).

Moreover, because ETTC is an extension of the conventional TTC and its definition is the same as the conventional TTC, so the TTC threshold can be also used for ETTC analysis. The TTC threshold is widely employed to identify risky situations. Vehicles with a TTC value lower than the threshold are considered as involving in collisions with high probability. Previous studies have suggested the TTC threshold vary from 1 to 3 s (Li et al., 2017a, 2014; Meng and Qu, 2012). Considering that there might have difference between the safety analysis results when the TTC threshold is set as different values, in this study, the ETTC threshold as 1 s, 2 s and 3 s respectively to validate the model results and reduce errors caused by different ETTC threshold settings.

3.3. Model formulation

Based on the vehicles' dynamic information obtained from the vehicle trajectory data and the definition of ETTC, we can calculate the ETTC of each vehicle between itself and its leading vehicle, and divide the following vehicle into two cases: (i) "Has the potential collision", if its ETTC values are less than ETTC threshold; (ii) "Doesn't have the potential collision", otherwise. According to these two cases, we can determine whether the dependent variable of models: y = 1 (Has the potential collision) and y = 0 (Doesn't have the potential collision). The following candidate variables in Table 1 are investigated that may affect the vehicle collision risk at toll plaza diverging area. Note that, the preliminary tests of correlations are applied for all independent variables and the italic ones are deleted after the test.

In this study, the total sample size is 75,732. Corresponding to different ETTC thresholds, the different percentages of dangerous sample are 4.39%, 9.82%, 15.67% respectively. Note that, the conflict of each vehicle is calculated every 0.1 s (counting by 3 frames) in this study. Each dataset is randomly divided into an original training dataset (including 80% of data) and a prediction dataset (including 20% of data). Further, for non-parametric models, the original training datasets were divided into a training subset and a validation subset using 10-fold cross validation method, which will be introduced in the following subsection

Table 1
Candidate variables and variables' explanations.

Category	Variable	Variable Explanation
Variables of the following vehicle	F_x	The x coordinates of the following vehicle in the ground coordinate, m.
	F_y	The y coordinates of the following vehicle in the ground coordinate, m.
	FTC_{type}	An indicator variable for the toll collection types of the following vehicle, 0 for a MTC vehicle, 1 for an ETC vehicle.
	$FL_{initial}$	The initial lanes of the following vehicle when it enters the toll plaza diverging area. (as shown in Fig. 2, 1 to 4)
	FL_{target}	*The following vehicle target toll collection lanes when it exits the toll plaza diverging area. (as shown in Fig. 2, 1 to 12)*
	F_v	The speed of the following vehicle, m/s.
	F_a	The angle between the following vehicle speed vector and X axis (as shown in Fig.4, Theta angle)
	T	*Time after the following vehicle entering the diverging area*
Variables of the leading vehicle	*Lx*	*The x coordinates of the leading vehicle in the ground coordinate, m.*
	Ly	*The y coordinates of the leading vehicle in the ground coordinate, m.*
	LTC_{type}	An indicator variable for the toll collection types of the leading vehicle, 0 for a MTC vehicle, 1 for an ETC vehicle.
	$LL_{initial}$	The initial lanes of the leading vehicle when it enters the toll plaza diverging area. (as shown in Fig. 2, 1 to 4)
	LL_{target}	*The leading vehicle target toll collection lanes when it exits the toll plaza diverging area. (as shown in Fig. 2, 1 to 12*
	L_v	The speed of the leading vehicle, m/s.
	L_a	The angle between the leading vehicle speed vector and X axis (as shown in Fig.4, Theta angle)
Other variables	Dij	The distance between two vehicles' centroids
	N	The number of total vehicles when the following vehicle passing the area
	N_{ETC}	*The number of ETC vehicles when the following vehicle passing the area*
	NMTC	*The number of MTC vehicles when the following vehicle passing the area*

Note: The italic ones are deleted after preliminary test of variable correlations.

4. Results and discussion

4.1. Results of LR models

Based on the vehicle trajectory data that was extracted from the videos, the LR model is applied first to estimate the relationship between crash risk and various factors and predict crash probability. The Pearson correlation checking of all the independent variables was adopted before estimating the LR model. The result shows that seven variables including FL_{target}, T, L_x, L_y, LL_{target}, $LL_{initial}$, N_{ETC} and N_{MTC} have significant and high correlations with other independent variables ($P < 0.05, |r| > 0.7$), and these variables were excluded in the model to avoid the error accumulation effects.

The model results and model performance of training and prediction are summarized in Table 2 and Table 4, respectively. Table 2 provides the final estimated results of LR models, variables that do not have significant impacts on collision risks are excluded in the final models, and the sign of coefficients are the same. Due to the datasets of dependent variable in these three models are different, so the significant independent variables in the results maybe also different. Table 4 shows that all the three LR models provide good model performance and the training and prediction accuracy are greater than 90%. Note that, the original training dataset was not further divided for validation, because the 10-fold cross validation is not applied for LR models.

Results show negative associations between Fx and collision risk which indicates vehicles become more dangerous when it is closer to toll collection lanes. This result is in accord with our previous study (Xing et al., 2019). The result of F_y also shows that vehicles are riskier when drive on the left side (the left side when driving toward the toll plaza) of toll plaza diverging area. The result is reasonable since more vehicles are willing to go through the toll collection lanes located on the left side of road, and the traffic on this side is more complicated and has higher traffic volume. On the other hand, because ETC lanes are usually located on the left side, almost all ETC vehicles drive on the leftward lanes. Moreover, according to the value of the coefficient for the FTC_{type} and LTC_{type}, it could be concluded that vehicles with ETC toll collection type have higher collision risks than MTC vehicles. Note that, the vertical position of the following vehicle (F_y) is significant only for the dataset when the ETTC threshold is 3 s. It might because that there are more dangerous samples when ETTC threshold is larger, which could influence the estimation results of model.

Furthermore, compared to the following vehicle with initial lane 1, the following vehicle whose initial lane is lane 2 are most easily to be involved in collisions, followed by initial lane 3 and initial lane 4. The finding is intuitively reasonable since vehicles come from initial lane 2 are generally locate in the middle of the longitudinal direction at toll

Table 2
Results of LR models.

Variables	ETTC threshold = 1 s		ETTC threshold = 2 s		ETTC threshold = 3 s	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
F_x	−0.001[a]	< .0001	−0.002[a]	< .0001	−0.002[a]	< .0001
F_y	—	—	—	—	−0.004[a]	0.003
$FL_{initial}$ (lane2 VS lane1)	0.698[a]	0.096	0.323[a]	0.069	0.318[a]	0.060
$FL_{initial}$ (lane3 VS lane1)	0.514[a]	0.091	0.257[a]	0.066	0.300[a]	0.057
$FL_{initial}$ (lane4 VS lane1)	0.409[a]	0.113	0.237[a]	0.081	0.256[a]	0.070
FTC_{type}	0.059[a]	0.060	0.025[a]	0.045	0.008[a]	0.039
F_v	0.305[a]	0.009	0.430[a]	0.007	0.527[a]	0.007
F_a	—	—	−0.038[a]	0.009	−0.053[a]	0.007
LTC_{type}	0.112[a]	0.057	—	—	0.041[a]	0.035
L_v	−0.289[a]	0.009	−0.469[a]	0.008	−0.565[a]	0.007
L_a	—	—	0.018[a]	0.009	0.008[a]	0.007
Dij	−0.554[a]	0.011	−0.330[a]	0.005	−0.239[a]	0.003
N	—	—	0.042[a]	0.017	0.072[a]	0.015
Constant	1.009[a]	0.278	1.335[a]	0.200	0.941[a]	0.168

[a] Significant at the 95% confidence level.

Table 3
Statistics of initial lanes.

		Initial 1	Initial 2	Initial 3	Initial 4	Total sample
MTC VehicleETC Vehicle		27	132	252	174	585
Total sample		94	180	126	31	431
		121	312	378	205	1016
ETTC threshold = 1 s	Safe sample	9254	21757	27058	14337	72406
	Dangerous sample	484	1157	1153	532	3326
	Total sample	9738	22914	28211	14869	75,732
ETTC threshold = 2 s	Safe sample	8561	20428	25815	13492	68296
	Dangerous sample	1177	2486	2396	1377	7436
	Total sample	9738	22914	28211	14869	75,732
ETTC threshold = 3 s	Safe sample	7910	18930	24429	12593	63862
	Dangerous sample	1828	3984	3782	2276	11870
	Total sample	9738	22914	28211	14869	75,732

Table 4
The accuracy of training, validation and prediction of different models.

Accuracy (%)	LR	KNN	SVM	ANN	DT	RF
			ETTC threshold=1 s			
Training	96.5	97.37	96.3	95.69	97.46	99.88
Validation	----------	95.15	95.72	95.56	96.48	95.82
Prediction	96.23	95.65	95.74	95.74	94.18	96.33
			ETTC threshold=2 s			
Training	93.6	95.33	94.48	92.89	94.9	99.85
Validation	----------	90.97	92.86	92.9	93.76	93.94
Prediction	93.82	90.43	90.72	91.02	92.46	92.08
			ETTC threshold=3 s			
Training	91.6	93.67	93.33	90.64	95.5	99.75
Validation	----------	87.42	90.98	90.51	91.78	92.04
Prediction	91.75	85.64	87.21	88.24	87.26	88.07

Note: The background color of the table represents the general trend of prediction accuracy. From red color to yellow and to green ones, the accuracy decreases accordingly. Red one represents the result with the lowest accuracy and green one represents the result with the highest accuracy.

plaza diverging area, which are more easily collide by surrounding vehicles. Table 3 shows the statistics of initial lanes. For vehicles comes from initial lane 1, up to 78% (94/121, as shown in Table 3) of them are ETC vehicles which don't need to change lanes for toll collection, so they are safer than vehicles form other initial lanes. Note that, although the result of F_y and $FL_{initial}$ both evaluate influences of vertical position, they are essentially different. The $FL_{initial}$ only evaluates the impact of initial vertical position on vehicle collision risk, while F_y estimates the distribution of collision risk in the whole diverging area. Thus, the estimated results are not necessarily consistent of these two variables.

The results of F_v and L_v indicate that higher speed of following vehicles could lead to higher collision risks, while higher speed of leading vehicle could decrease collision risks. A possible explanation is that the following vehicle with higher speed need longer braking distance and the preceding vehicle with lower speed could not provide enough distance for the following vehicle, which could cause the higher collision risks. It is also supported by the estimated result of the distance between two vehicles' centroids (D_{ij}), which has the negative associations with collision risk, since the larger distance provides more space for the following vehicle to avoid crashes.

Finally, the angle between the following vehicle speed vector and X axis has a negative effect while that of the leading vehicle has a positive effect on the vehicle collision risks. As shown in the Fig.4, the angle between the two vehicles' speed vector will be larger when the Theta angle of the following vehicle is smaller or the Theta angle of the leading vehicle is larger. Therefore, the estimate results are reasonable because that larger angle between the two vehicles could maybe result in more serious collisions when the speed of two vehicles are constant.

4.2. Results of non-parametric models

In this study, we utilized five models to predict collision risk based on prediction subsets, including KNN, SVM, ANN, DT and RF models. In order to improve model performance, the 10-fold cross validation was adopted. As shown in Fig. 5, the original training dataset D (80% of the total dataset) is split into 10 subsets D_i, $i \in \{1, 2, ..., 10\}$. For each iteration, a D_i is selected as the validation subset and a result i is output. Then, the final outcome is calculated by averaging these results.

Besides 10-fold cross validation, another vitally important step is to select key parameter values. In the present study, we utilized accuracy curves of training and validation processes to determine the best one. These curves with different ETTC thresholds are displayed in Figs. 6–8. As shown in Fig. 6, all five non-parametric models present various accuracy curves. Generally, a high training accuracy indicates a good

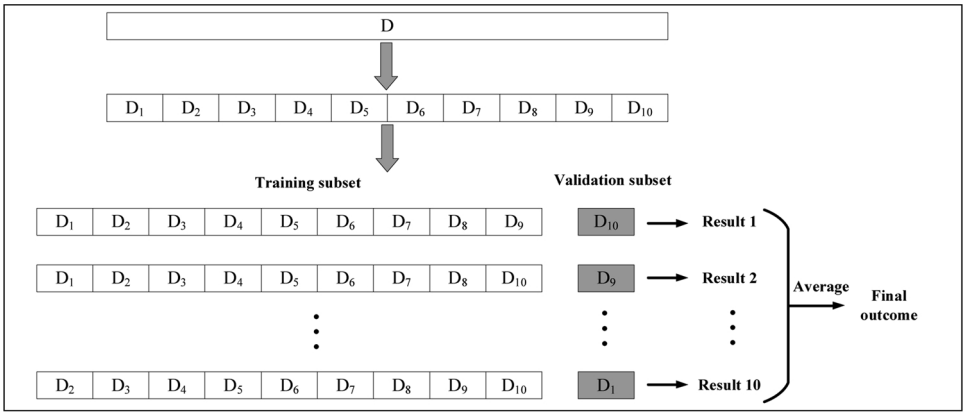


Fig. 5. Illustration of 10-fold cross validation.

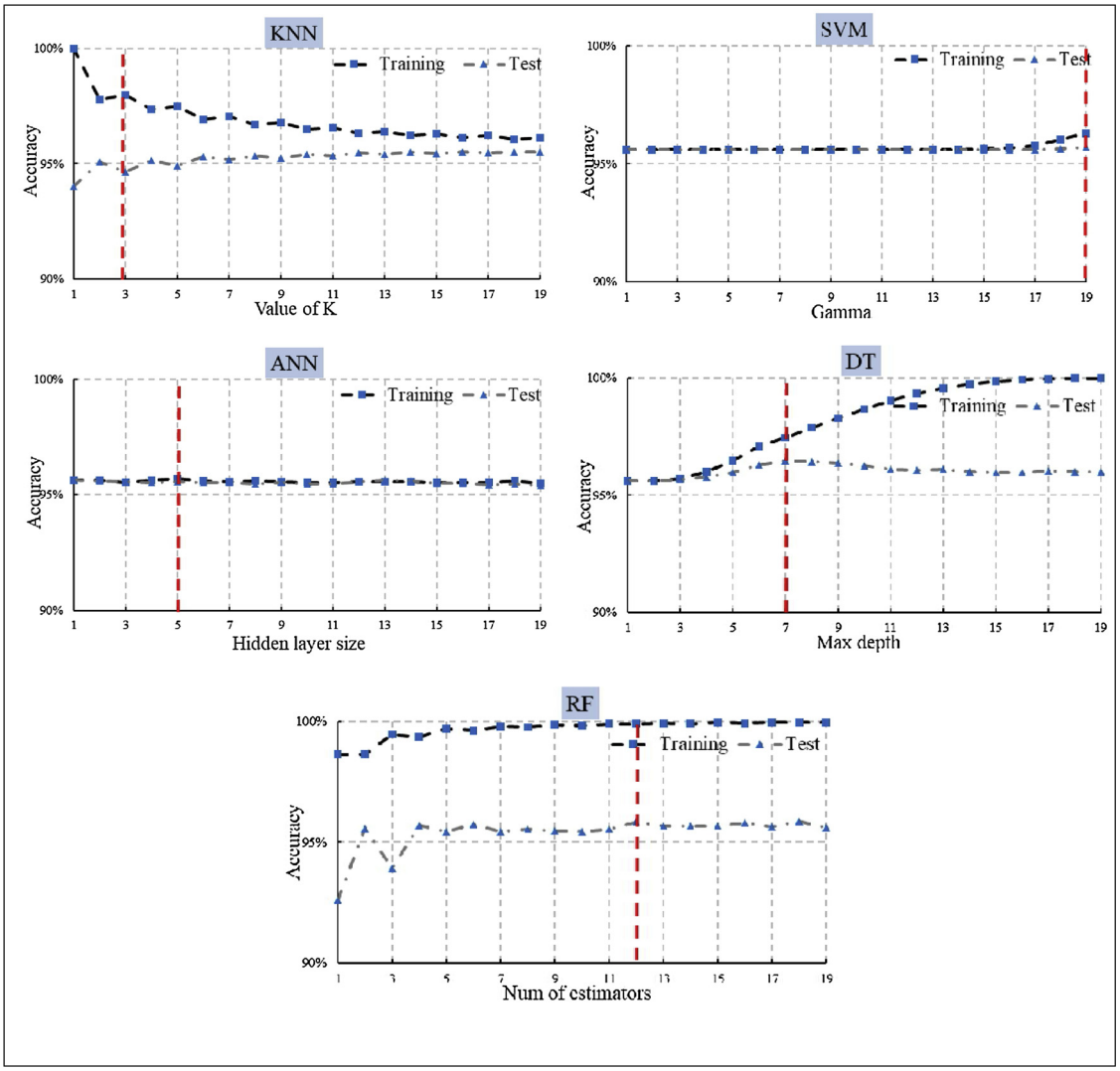


Fig. 6. Parameter value selection of different models (ETTC threshold = 1 s).

prediction performance for the training subset, while a high validation accuracy can avoid the over-fitting phenomenon. Thus, under the premise of a good validation accuracy, we select the key parameter values with a high accuracy of training process.

According to Fig. 6, the training accuracy curve of KNN shows a fluctuating downward, while the change of validation accuracy is not significant. The $K = 3$ is selected on the basis of the highest validation accuracy as well as a relatively high training one. With respect to SVM, the parameter gamma is set as a geometrical sequence with 19 different values between 10-6 and 10-2.3. These 19 values are tested as the candidates of gamma and the curves display a different trend. With a stable accuracy before the seventh gamma value, the training curve continues to increase remarkably. The validation curve, however, displays an arch shape with the highest accuracy at the fourteenth value of gamma, which is chosen as the best one in our study. The training and validation curves of ANN show the same trend, and thus the value corresponding to highest accuracy is selected. The curves of DT are similar to those of SVM. The maximum depth of 8 is determined, in which the validation accuracy is highest and training accuracy is also good enough. Regarding RF model, both curves increase gradually. In order to avoid over-fitting, the value of 9 is chosen in this study. Results of TTC threshold of 2 s (in Fig. 7) and 3 s (in Fig. 8) are similar to those in Fig. 6.

With the determined key parameter values, we utilized these five models to predict collision risk based on prediction subsets, the model performance is shown and discussed in the next subsection. An example of DT is shown in Fig. 9. It indicates that following vehicle's speed, leading vehicle's speed and the distance between two vehicles' centroids are the main splitters in the decision tree. These three variables play an important role in classifying vehicle collision risk. The tree was first split by the distance between two vehicles' centroids, it indicates that it is the most important variable to classify and predict collision risk and the specific numerical classification shows that it has the negative effects on vehicle collision risks. Furthermore, the effects of leading vehicle's speed and following vehicle's speed on crash risk are also consistent with the results in sub Section 4.1.

4.3. Model performance comparison

In this study, the model accuracy is applied to compare model performance, it is defined as the percentage of correct prediction. Generally, the models with larger accuracy outperform the models with smaller accuracy. Model performance results of model training, validation, as well as prediction are displayed in Table 4.

In Table 4, most of these models have a model accuracy higher than 90%. The comparison results shown that KNN, SVM, DT and RF models have better model performance than LR model in model training, while the ANN model has the worst model performance. However, in model

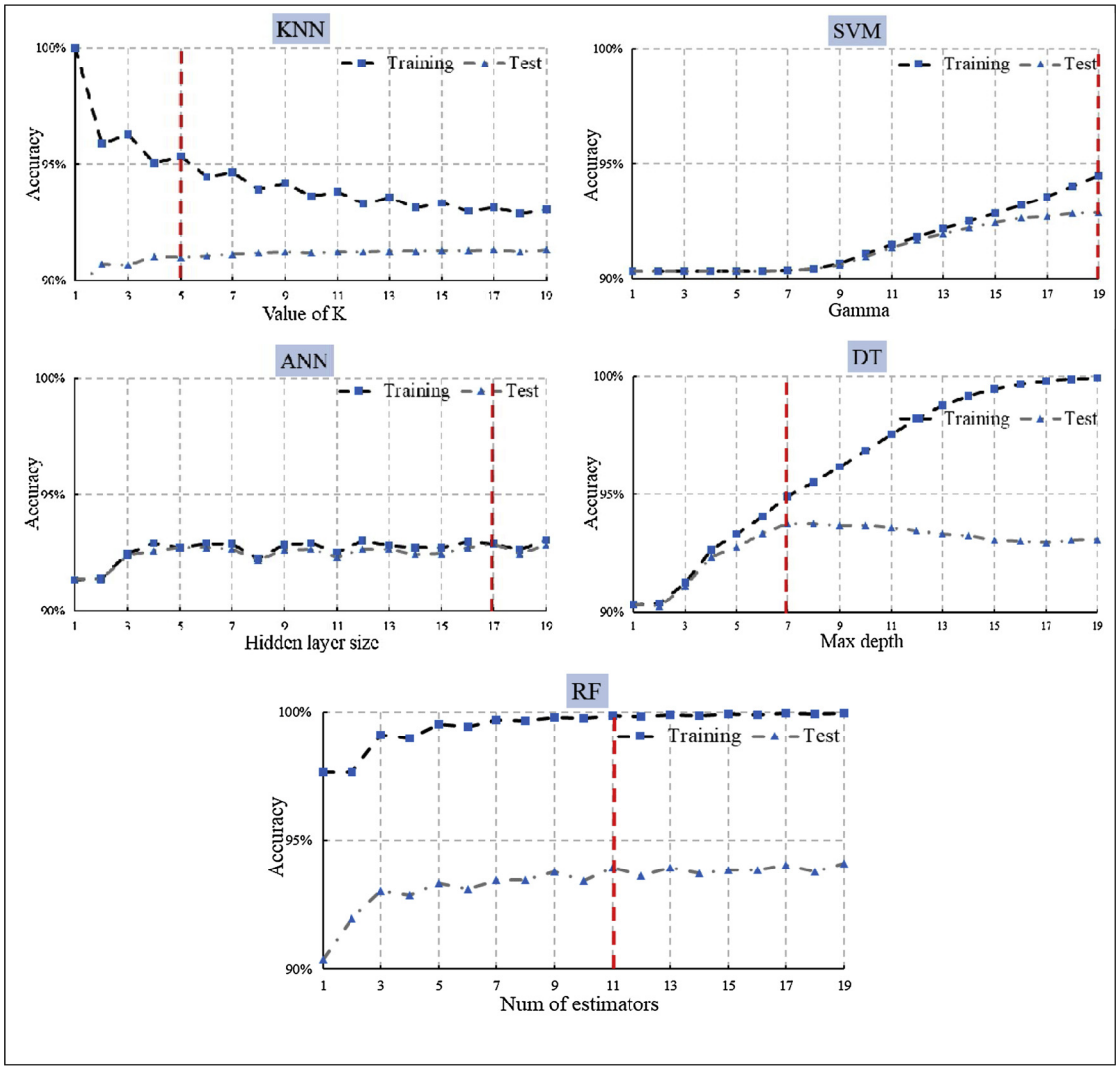


Fig. 7. Parameter value selection of different models (ETTC threshold = 2 s).

prediction, the accuracy of LR model is higher than that of other five non-parametric models under various ETTC thresholds conditions. In addition, the lager value of ETTC threshold, the lower accuracies of models have. When the value of ETTC threshold is 3 s, except the LR model, the accuracies of other five non-parametric models are all lower than 90%.

The results of the LR model and five non-parametric models imply several interesting findings. First, collision risk can be quantitatively evaluated based on different models with good performance utilizing trajectory data. Most of these models have a prediction accuracy higher than 90%. The results indicate the effectiveness of these models and usability of trajectory data. The trajectory data include a large number of dynamic information of each vehicle, and thus provide more details for collision risk analysis.

Second, not all non-parametric models have a better model performance than the LR model. Most non-parametric models perform better in the training datasets but worse in the prediction ones. The possible reason is that the non-parametric models cannot avoid the overfitting phenomenon due to the complex model structure. The LR model, however, perform better for prediction datasets since its form is simple and effective. Another possible explanation of this result is the special data structure of microscopic vehicle trajectory. Previously, some studies shown the good performance of non-parametric models for safety analysis using macro-level data (Delen et al., 2006; Huang et al., 2018; Li et al., 2012). For example, Dong et al., 2015 utilized SVM model for crash prediction based on crash data as well as traffic volume, speed, daily vehicle miles traveled and so on. All the data in these prior studies are aggregated in the time interval of 15 min, 30 min and even 1 h. The structure of these macroscopic data is definitely different from that of the microscopic trajectory data used in this study. The trajectory data is collected at each 0.1 s and more dynamic characteristics of vehicle operations are included. The microscopic dependent variable is also the conflict instead of crash, and these differences may result in the different model performances.

Additionally, the different values of ETTC threshold result in different datasets of dependent variable, which have different proportion of dangerous samples (4.39%, 9.82%, and 15.67% respectively). It is essential reason for the different model performances. Furthermore, there are significant difference of accuracies among three datasets for each non-parametric model. However, the difference is relatively small for the LR model.

Finally, even for the training datasets, although the non-parametric models have relatively good performances as they could avoid the inherent problems and provide high accuracy, they may not be the best choice for safety analysis, since the results could be difficult to be interpreted and may not be helpful for understanding the relationship between crash risks and various contributing factors (Weng and Meng, 2011). Taking the DT model shown in Fig. 9 as an example. Although

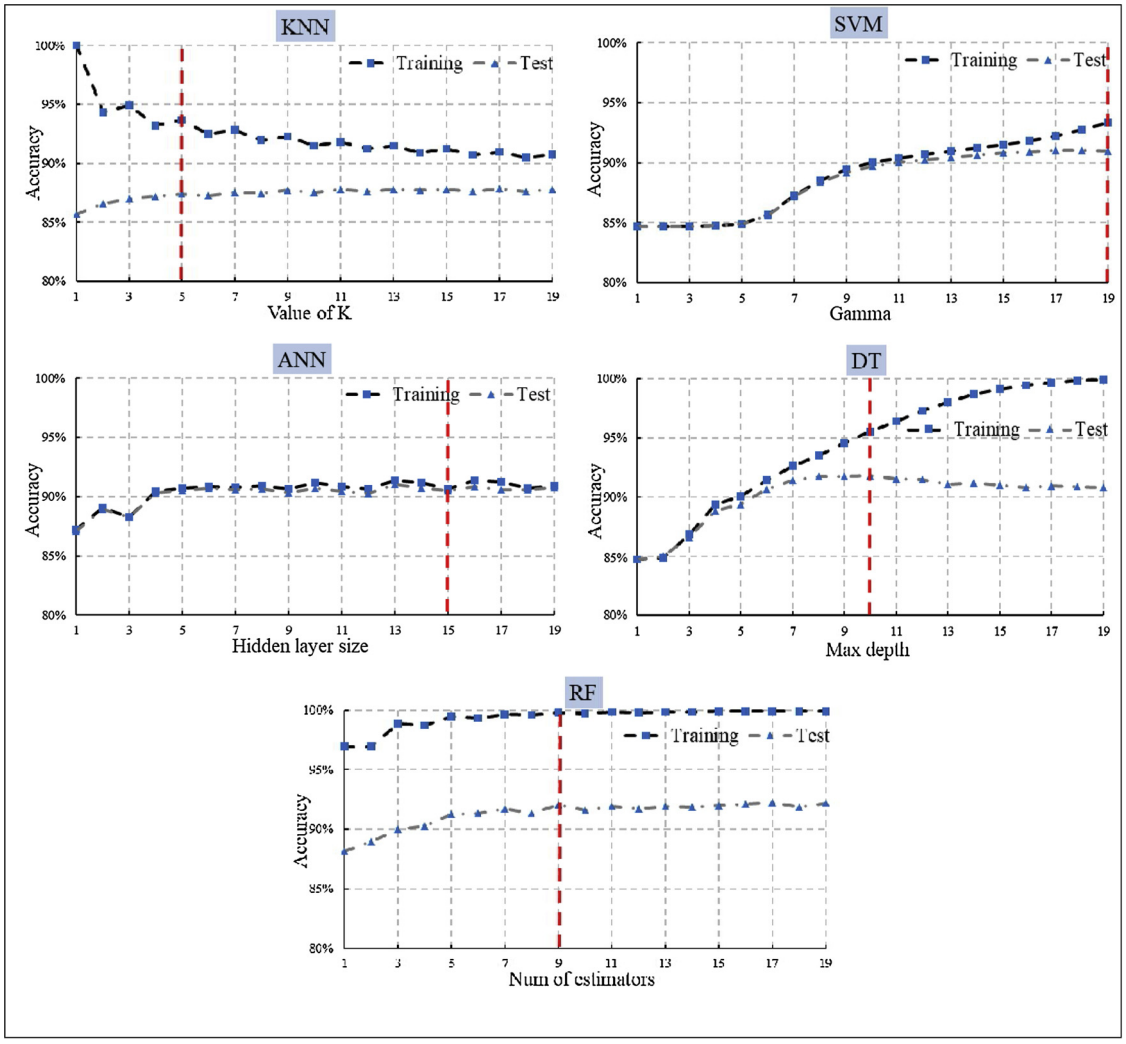


Fig. 8. Parameter value selection of different models (ETTC threshold = 3 s).

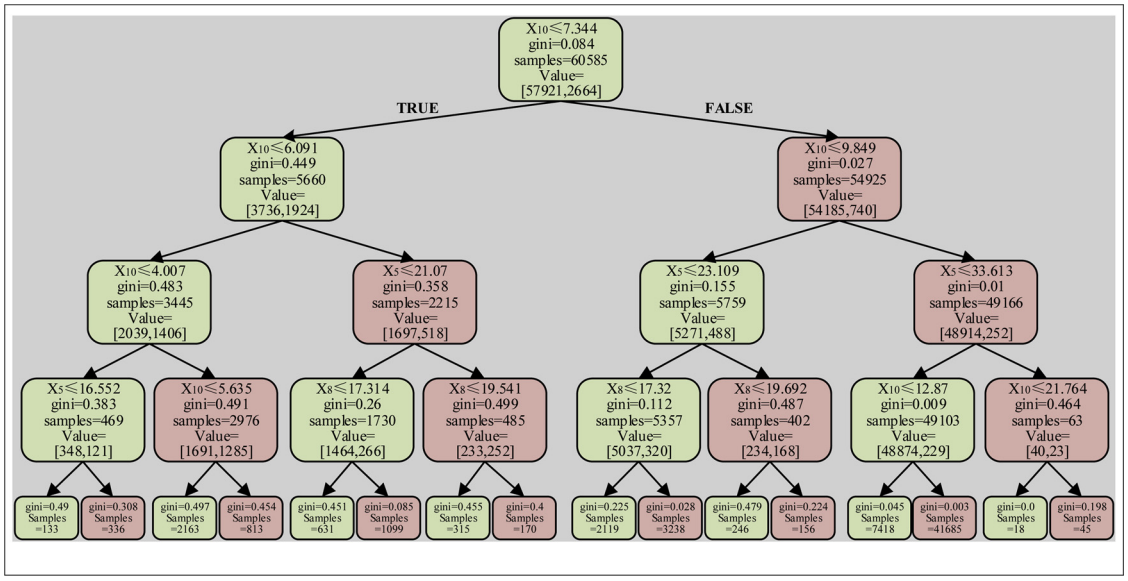


Fig. 9. Result of DT with a max depth of 4 (ETTC threshold = 1 s).
Note: X5 is Fv, the speed of following vehicle, m/s; X8 is Lv, the speed of the leading vehicle, m/s; X10 is Dij, the distance between two vehicles' centroids.

this model has a model accuracy of higher than 90%, the structure of the decision tree is complicated. If there is a larger magnitude of datasets, the complexity could be greater. Moreover, CART always yields binary trees, which only can divide variables into two categories and cannot analyze the impacts of all variables on collision risk. In the case of vehicle crashes, it may not be very practical when it is used to evaluate the effect of a specific variable on collision risk (Breiman et al., 1984; de Oña et al., 2013).

The results show that LR model always have good model performance, however, it should be admitted that the LR modeling approach also has advantages and disadvantages. On one hand, LR model is more suitable for predicting the relationships between various independent variables and a discrete dependent variable. It can be applied to predict the probability of the event and estimate the marginal effect of each independent variable. On the other hand, the incorrect estimations of crash risk likelihood are easy to occur if there is the violation of model assumption, and it is hard to eliminate the error caused by variables' correlation completely. More importantly, the LR modeling approach cannot to compare the effects of various factors on collision risk and it also cannot highlight sophisticated relationships that are difficult to discover.

For these reasons, a few previous studies suggested that the complementary use of both modeling approach is a promising approach to perform the difficult task (Muller and Möckel, 2008; Shi and Abdel-Aty, 2015). Meanwhile, there are plenty of studies attempted to look for complementary approaches for limitation elimination and verified the availability of them. For example, the "elasticity effect" is a better approach to estimate the effect of each contributing factor and compare the effects of different independent variables on collision risk. This approach is usually employed after the LR model, and can calculate the changes of collision probability caused by a change in the independent variable, which can provide a good further explanation of LR modeling results. (Eluru et al., 2008; Guo et al., 2018; Xing et al., 2019). Therefore, considering the requirement for adequate information of model results, the higher model accuracy and the existing solutions for modeling limitation elimination, the LR model is still a good choice for our study.

5. Conclusions and recommendation

This study aimed to suggest the best model for vehicle safety evaluation at toll plaza diverging area. Vehicle trajectory data were extracted by computer vision techniques using UAV videos and ETTC indicator was used to evaluate the collision risk of unconstrained vehicle motions. Three different values of TTC threshold are set for model validation. Moreover, the LR model and five typical non-parametric models including KNN, SVM, ANN, DT and RF were developed for model estimation, the model performance were compared. The important findings are expressed as follows:

(1) Collision risks could be quantitatively evaluated based on different models (both non-parametric models and LR model) based on trajectory data with good prediction performance;
(2) Not all non-parametric models have a better model performance than the LR model, and most non-parametric models perform better in the training datasets but worse in the prediction ones. The KNN, SVM, DT and RF models have better model performance than LR model in model training, and the ANN model has the worst model performance. However, in model prediction, the accuracy of LR model is higher than that of other five non-parametric models under various ETTC thresholds conditions.
(3) Even for the non-parametric models have relatively good model performances in model training, the weakness of model result interpretation limits the practicability of the non-parametric models. Considering the LR model also has good prediction accuracy and there are existing solutions to eliminate the modeling limitations, so in the perspective of the practical application, the LR model is still a good choice for collision risk evaluation using trajectory data at the toll plaza diverging area.
(4) In addition, the different datasets of dependent variable caused by the different values of ETTC threshold could have significant effects on model results, especially for non-parametric models. However, the difference of accuracies among three datasets is relatively small for the LR model. It is suggested to take two or more values of ETTC threshold for model result validation in the safety analysis.
(5) The results of LR model indicate that there are significant different effects on collision risk among vehicles with different factors. More specifically, vehicles higher collision risk when they drive on the left side of toll plaza diverging area and more dangerous situations could be found for an ETC vehicle. Further, compared to the vehicles coming from initial lane 1, the following vehicle whose initial lane is lane 2 are most easily to be involved in collisions, followed by initial lane 3 and initial lane 4. In addition, the vehicle collision risks are positively associated with the speed of the following vehicle and the angle between the leading vehicle speed vector and X axis. On the other hand, negative associations were found between the vehicle collision risk and other factors, such as the speed of the following vehicle, the distance between two vehicles, the angle between the following vehicle speed vector and X axis. Furthermore, the results of DT model suggest that following vehicle's speed, leading vehicle's speed and the distance between two vehicles' centroids play an important role in classifying vehicle collision risk. The effects of them on crash risk are consistent with the results of LR model.

These findings could provide more helpful information to the safety evaluation and improvement of the toll plaza diverging area. However, the authors recognize that much still needs to be done in the next step. The detailed influence needs to be investigated in order to further enhance prediction accuracy. Besides, the interpretability and convenience of non-parametric models could be improved, which may impede the practicability compared with statistical models. And the unobserved heterogeneity should be analysis by employing the advanced modeling techniques. In addition, the results may only be applicable to this dataset. Further studies of various toll plaza diverging area should be conducted for validation and a general model suitable for all kinds of toll plaza diverging area should be developed.

Declaration of Competing Interest

There are no interests to declare.

Acknowledgements

The author would like to thank the Fundamental Research Funds for the Central Universities and the Postgraduate Research & Practice Innovation Program of Jiangsu Province [KYCX17-0148]. Moreover, thanks to the National Natural Science Foundation of China [Grant No.51778141 and No.71601046]. Part of the research was conducted at the University of Central Florida where the first author spent a year as a visiting student funded by China Scholarship Council.

References

Abdel-Aty, M.A., Hassan, H.M., Ahmed, M., Al-Ghamdi, A.S., 2012. Real-time prediction of visibility related crashes. Transp. Res. Part C Emerg. Technol. 24, 288–298. https://doi.org/10.1016/j.trc.2012.04.001.

Abdelwahab, H., Abdel-Aty, M., 2002. Artificial neural networks and logit models for traffic safety analysis of toll plazas. Transp. Res. Rec. J. Transp. Res. Board 1784, 115–125. https://doi.org/10.3141/1784-15.

Abdelwahab, H.T., Abdel-Aty, M.A., 2007. Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. Transp. Res. Rec. J. Transp. Res. Board 1746 (1), 6–13. https://doi.org/10.3141/1746-02.

Abellán, J., López, G., De Oña, J., 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Syst. Appl. 40 (15), 6047–6054. https://doi.org/10.1016/j.eswa.2013.05.027.

Abuzwidah, M., Abdel-Aty, M., 2015. Safety assessment of the conversion of toll plazas to all-electronic toll collection system. Accid. Anal. Prev. 80, 153–161. https://doi.org/10.1016/j.aap.2015.03.039.

Abuzwidah, M.A.M., 2011. Evaluation and Modeling of the Safety of Open Road Tolling System. Masters Thesis September 2011, 103. .

Ali, E.M., Ahmed, M.M., Wulff, S.S., 2019. Detection of critical safety events on freeways in clear and rainy weather using SHRP2 naturalistic driving data: parametric and non-parametric techniques. Saf. Sci. January 2018, 1–9. https://doi.org/10.1016/j.ssci.2019.01.007.

Altman, N.S., Altman, N.S., 2012. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression 1305.

Behbahani, H., Nadimi, N., 2015. A framework for applying surrogate safety measures for sideswipe conflicts. Int. J. Traffic Transp. Eng. 5 (4), 371–4383. https://doi.org/10.7708/ijtte.2015.5(4).03.

Breiman, L.E.O., 2001. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees, Wadsworth & Brooks/Cole Advanced Books & Software. Pacific Grove, CA.

Carroll, K., 2016. Evaluation of Real World Toll Plazas Using Driving Simulation.

Chang, L.Y., Chien, J.T., 2013. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Saf. Sci. 51 (1), 17–22. https://doi.org/10.1016/j.ssci.2012.06.017.

Cheng, L., Chen, X., Vos, J., De Lai, X., Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel Behav. Soc. 14 (August), 1–10. https://doi.org/10.1016/j.tbs.2018.09.002. 2018.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Das, A., Abdel-Aty, M., Pande, A., 2009. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. J. Safety Res. 40 (4), 317–327. https://doi.org/10.1016/j.jsr.2009.05.003.

de Oña, J., López, G., Abellán, J., 2013. Extracting decision rules from police accident reports through decision trees. Accid. Anal. Prev. 50, 1151–1160. https://doi.org/10.1016/j.aap.2012.09.006.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38 (3), 434–444. https://doi.org/10.1016/j.aap.2005.06.024.

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. Accid. Anal. Prev. 82, 192–198. https://doi.org/10.1016/j.aap.2015.05.018.

Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accid. Anal. Prev. 40 (3), 1033–1054. https://doi.org/10.1016/j.aap.2007.11.010.

Guo, Y., Osama, A., Sayed, T., 2018. A cross-comparison of different techniques for modeling macro-level cyclist crashes. Accid. Anal. Prev. 113, 38–46. https://doi.org/10.1016/j.aap.2018.01.015.

Guo, Y., Sayed, T., Zaki, M.H., Liu, P., 2016. Safety evaluation of unconventional outside left-turn lane using automated traffic conflict techniques. Am. J. Civ. Eng. 43 (7), 631–642. https://doi.org/10.1139/cjce-2015-0478.

Hayward, J.C., 1972. Near-miss Determination Through Use of a Scale of Danger. Highw. Res. Rec. doi:TTSC 7115.

Huang, H., Peng, Y., Wang, J., Luo, Q., Li, X., 2018. Interactive risk analysis on crash injury severity at a mountainous freeway with tunnel groups in China. Accid. Anal. Prev. 111 (July), 56–62. https://doi.org/10.1016/j.aap.2017.11.024. 2017.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. 108 (September), 27–36. https://doi.org/10.1016/j.aap.2017.08.008.

Jung, S., Qin, X., Oh, C., 2016. Improving strategic policies for pedestrian safety enhancement using classification tree modeling. Transp. Res. Part A Policy Pract. 85, 53–64. https://doi.org/10.1016/j.tra.2016.01.002.

Kecman, V., 2005. Support vector machines – an introduction. In Support Vector Machines: Theory and Applications. Springer, Berlin, Heidelberg, pp. 1–47.

Kuhnert, P.M., Do, K.A., McClure, R., 2000. Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Comput. Stat. Data Anal. 34 (3), 371–386. https://doi.org/10.1016/S0167-9473(99)00099-7.

Laureshyn, A., Svensson, Å., Hydén, C., 2010. Evaluation of traffic safety, based on micro-level behavioural data: theoretical framework and first implementation. Accid. Anal. Prev. 42 (6), 1637–1646. https://doi.org/10.1016/j.aap.2010.03.021.

Li, Y., Li, Z., Wang, H., Wang, W., Xing, L., 2017a. Evaluating the safety impact of adaptive cruise control in traffic oscillations on freeways. Accid. Anal. Prev. 104, 137–145. https://doi.org/10.1016/j.aap.2017.04.025.

Li, Y., Wang, H., Wang, W., Liu, S., Xiang, Y., 2016. Reducing the risk of rear-end collisions with infrastructure-to-vehicle (I2V) integration of variable speed limit control and adaptive cruise control system. Traffic Inj. Prev. 17 (6), 597–603. https://doi.org/10.1080/15389588.2015.1121384.

Li, Y., Xu, C., Xing, L., Wang, W., 2017b. Integrated cooperative adaptive cruise and variable speed limit controls for reducing rear-end collision risks near freeway bottlenecks based on micro-simulations. IEEE trans. Intell. Transp. Syst. 18 (11), 3157–3167. https://doi.org/10.1109/TITS.2017.2682193.

Li, Z., Li, Y., Liu, P., Wang, W., Xu, C., 2014. Development of a variable speed limit strategy to reduce secondary collision risks during inclement weathers. Accid. Anal. Prev. 72, 134–145. https://doi.org/10.1016/j.aap.2014.06.018.

Li, Z., Liu, P., Wang, W., Xu, C., 2012. Using support vector machine models for crash injury severity analysis. Accid. Anal. Prev. 45, 478–486. https://doi.org/10.1016/j.aap.2011.08.016.

Mahmud, S.M.S., Ferreira, L., Hoque, M.S., Tavassoli, A., 2017. Application of proximal surrogate indicators for safety evaluation: a review of recent developments and research needs. IATSS Res. 41 (4), 153–163. https://doi.org/10.1016/j.iatssr.2017.02.001.

Mckinnon, I.A., 2013. Operational and Safety-based Analyses of Varied Toll Lane Configurations.

Meng, Q., Qu, X., 2012. Estimation of rear-end vehicle crash frequencies in urban road tunnels. Accid. Anal. Prev. 48, 254–263. https://doi.org/10.1016/j.aap.2012.01.025.

Meng, Q., Weng, J., 2011. Evaluation of rear-end crash risk at work zone using work zone traffic data. Accid. Anal. Prev. 43 (4), 1291–1300. https://doi.org/10.1016/j.aap.2011.01.011.

Muller, R., Möckel, M., 2008. Logistic regression and CART in the analysis of multimarker studies. Clin. Chim. Acta 394 (1–2), 1–6. https://doi.org/10.1016/j.cca.2008.04.007.

Mussone, L., Ferrari, A., Oneta, M., 1999. An analysis of urban collisions using an artificial intelligence model. Accid. Anal. Prev. 31 (6), 705–718. https://doi.org/10.1016/S0001-4575(99)00031-7.

Pakgohar, A., Tabrizi, R.S., Khalili, M., Esmaeili, A., 2011. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach. Procedia Comput. Sci. 3, 764–769. https://doi.org/10.1016/j.procs.2010.12.126.

Perkins, S.R., Harris, J.L., 1968. Traffic conflict characteristics: accident potential at intersections. Highw. Res. Board Rec. 225 (225), 35–44. https://doi.org/10.2307/1308182.

Porto-pazos, A.B., Veiguela, N., Mesejo, P., Navarrete, M., Alvarellos, A., 2011. Arti. Astrocytes Improve Neural Network Perform. 6 (4), 1–8. https://doi.org/10.1371/journal.pone.0019109.

Saad, M., Abdel-Aty, M., Lee, J., 2018. Analysis of driving behavior at expressway toll plazas. Transp. Res. Part F Traffic Psychol. Behav. 61, 163–177. https://doi.org/10.1016/j.trf.2017.12.008.

Sayed, T., Zaki, M.H., Autey, J., 2013. Automated safety diagnosis of vehicle-bicycle interactions using computer vision analysis. Saf. Sci. 59, 163–172. https://doi.org/10.1016/j.ssci.2013.05.009.

Sayed, T., Zein, S., 1999. Traffic conflict standards for intersections. Transp. Plan. Technol. 22 (4), 309–323. https://doi.org/10.1080/03081069908717634.

Shi, Q., Abdel-Aty, M., 2015. Big data applications in real time traffic operation and safety monitoring and improvement on urban expressways.pDf. Transp. Res. Part C Emerg. Technol. 58, 380–394.

Siddiqui, C., Abdel-Aty, M., Huang, H., 2012. Aggregate nonparametric safety analysis of traffic zones. Accid. Anal. Prev. 45, 317–325. https://doi.org/10.1016/j.aap.2011.07.019.

Wang, E.G., Sun, J., Jiang, S., Li, F., 2017. Modeling the various merging behaviors at expressway on-ramp bottlenecks using support vector machine models. Transp. Res. Procedia 25, 1327–1341. https://doi.org/10.1016/j.trpro.2017.05.157.

Ward, J.R., Agamennoni, G., Worrall, S., Bender, A., Nebot, E., 2015. Extending Time to Collision for probabilistic reasoning in general traffic scenarios. Transp. Res. Part C Emerg. Technol. 51, 66–82. https://doi.org/10.1016/j.trc.2014.11.002.

Weng, J., Meng, Q., 2012. Effects of environment, vehicle and driver characteristics on risky driving behavior at work zones. Saf. Sci. 50 (4), 1034–1042. https://doi.org/10.1016/j.ssci.2011.12.005.

Weng, J., Meng, Q., 2011. Analysis of driver casualty risk for different work zone types. Accid. Anal. Prev. 43 (5), 1811–1817. https://doi.org/10.1016/j.aap.2011.04.016.

Weng, J., Meng, Q., Yan, X., 2014. Analysis of work zone rear-end crash risk for different. Accid. Anal. Prev. 72, 449–457.

Wu, Y., Abdel-Aty, M., Cai, Q., Lee, J., Park, J., 2018. Developing an algorithm to assess the rear-end collision risk under fog conditions using real-time data. Transp. Res. Part C Emerg. Technol. 87, 11–25. https://doi.org/10.1016/j.trc.2017.12.012.

Wu, Y., Abdel-aty, M., Zheng, O., 2019. Developing a crash warning system for the Bike Lane Area at intersections with connected vehicle technology. Transp. Res. Rec. J. Transp. Res. Board 2673 (4), 1–12. https://doi.org/10.1177/0361198119840617.

Wu, Y., Abdel-Aty, M., Zheng, O., Cai, Q., Zhang, S., 2019b. Automated Safety Diagnosis Using Unmanned Aerial Vehicles Based on Deep Learning. Under riview.

Xing, L., He, J., Abdel-Aty, M., Cai, Q., Li, Y., Zheng, O., 2019. Examining traffic conflicts of up stream toll plaza area using vehicles' trajectory data. Accid. Anal. Prev. 125 (August), 174–187. https://doi.org/10.1016/j.aap.2019.01.034. 2018.

Xu, C., Liu, P., Wang, W., Li, Z., 2015. Safety performance of traffic phases and phase transitions in three phase traffic theory. Accid. Anal. Prev. 85, 45–57. https://doi.org/10.1016/j.aap.2015.08.018.

Xu, C., Wang, W., Liu, P., 2013. A genetic programming model for real-time crash prediction on freeways. IEEE trans. Intell. Transp. Syst. 14 (2), 574–586. https://doi.org/10.1109/TITS.2012.2226240.

Yuan, J., Abdel-Aty, M., 2018. Approach-level real-time crash risk analysis for signalized intersections. Accid. Anal. Prev. 119, 274–289.

Yuan, J., Abdel-Aty, M., Wang, L., Lee, J., Yu, R., Wang, X., 2018. Utilizing bluetooth and adaptive signal control data for real-time safety analysis on urban arterials. Transp. Res. Part C Emerg. Technol. 97, 114–127. https://doi.org/10.1016/J.TRC.2018.10.009.

Zheng, L., Ismail, K., Meng, X., 2014. Traffic conflict techniques for road safety analysis: open questions and some insights. Can. J. Civ. Eng. 41 (7), 633–641. https://doi.org/10.1139/cjce-2013-0558.